THIS WEEK

EDITORIALS

PHOTOSYNTHESIS Enzyme upgrade helps plants grow faster **p.280**

WORLD VIEW Research can bolster new UN world goals **p.281**

FOSSIL Ancient spiral-toothed shark went for soft centres **p.282**

Diversity challenge

There is growing evidence that embracing diversity — in all its senses — is key to doing good science. But there is still work to be done to ensure that inclusivity is the default, not the exception.

Earlier this year, Tom Welton, a chemistry professor at Imperial College London, wrote about the prejudice he experiences as a gay scientist. Intolerant peers jump to conclusions, insult him and make assumptions about his beliefs and behaviour. It is better, Welton wrote, to hide behind a lie: "I often find it easier to say 'I'm a teacher'."

Scientist colleagues had no problem with him being gay. But he found that people in the lesbian, gay, bisexual and transgender community seemed to have a problem with him being a scientist. "Most scientists, medics and engineers know that unless they have a stethoscope around their neck they aren't valued," he wrote.

As we explore in a Feature on page 297, others may have a different experience. Scientists, of course, should not be judged by their sexuality. The principles of research — reliance on data, rigorous experimentation and respect for evidence — do not cluster by any of the ways that humans choose to define themselves and each other. Gender, race, ethnic background, social status, wealth, nationality, age, skin colour and sexuality are as irrelevant to doing science as a person's musical taste or dietary preference. Or are they?

There is no place in science (or outside it) for prejudice. But there must be a place for diversity, and there is growing evidence that such variety is a key ingredient for doing good science. Much of that evidence is discussed this week in a joint special issue of *Nature* and our sister publication *Scientific American*.

Diversity is a vague word. The special-issue content (available at nature.com/diversity) is wide-ranging and covers much ground. It can be usefully tied together by a working definition: diversity means an inclusive approach, both to the science itself and the make-up of the groups of people who carry out the research.

Diversity is a topic too often discussed in the negative, through stories of discrimination and bias against select communities. Science has its problems here just like most of society, and *Nature* has long spoken out, for example, against the under-representation of women. Much of the special-issue content frames the subject in a different way, and examines the benefits of an inclusive approach.

Attention, busy scientists: if diversity sounds like a worthy topic but one better left to your university's human-resources department, then turn to page 305, where Richard Freeman and Wei Huang explain how it might boost your citation rate. Their analysis of the surnames of US-based authors on some 2.5 million research papers suggests that scientists who tend to stick with their own kind publish less-cited work, and in lower-impact journals.

Why published collaborations with a greater mixture of surnames perform better is unknown. What is clearer is that a mixture of people (mixed across whatever divisions you care to mention) will be able to consider and to enable a wider range of possible solutions to a problem. If the problem is scientific, then the result of that diversity can be better science. On page 301, for example, Esteban Burchard describes how his ethnic background and experiences with a variety of cultures have helped him to study the genetics of asthma in Latino Americans. On page 304, Mónica Ruiz-Casares highlights how the results of mental-health research based on adult, Western populations might not apply to other cultures and communities.

DIVERSITY
A *Nature* and *Scientific American* special issue **nature.com/diversity**

"There is a place for positive discrimination to address specific imbalances."

But collaboration that spans vast personal differences can raise problems. On page 303, Wenzel Geissler and Ferdinand Okwaro discuss the sometimes-fraught scenarios that arise when researchers from very different economic backgrounds work closely together. To draw attention to this inequality can be awkward, say the duo, but that is better than the destructive ways it can surface if ignored.

Science has already been through one revolution in diversity. Traditional academic silos that held subjects as distinct disciplines have crumbled. Interdisciplinary research now sets the agenda in many fields, especially those with a direct impact on society, such as climate-change research. That shift, although beneficial, was not entirely spontaneous. It was managed and encouraged by senior scientists and funders, who saw the pay-off. To fully develop the benefits of diversity, to ensure that science becomes fully inclusive, similar intervention is necessary — even if it is as simple as a busy lab head stopping to consider the issue for the first time.

As a telling graphic in the special issue of *Scientific American* illustrates, 51% of the science and engineering workforce in the United States is white and male. There is a place for positive discrimination to address specific imbalances. But diversity is not just a case of championing minority interests — the benefits of diversity go to the majority. ■

A worthy ambition

Finalizing the European Research Area is still a vibrant and relevant goal.

The completion of the European Research Area remains a "gradual process", admits the European Commission rather forlornly, at the conclusion of a report it published earlier this week on progress towards an entity within which European researchers and their ideas can circulate freely.

The European Research Area (ERA) was originally due to be finalized by the end of this year. The notion that this could happen, set

in train only two years ago by Máire Geoghegan-Quinn — who will depart as research commissioner of the European Union (EU) this autumn — was always as fanciful as it was beside the point.

That is because the ERA is a process, not an event. The project will never end. Anyone who imagines that it might do so only has to look at the United States. There, despite a genuinely single market and decades of federal incentives, huge disparities persist in 'research excellence' — however it is measured — between, say, Massachusetts and Montana.

That the problem is difficult does not mean that it should not be addressed. Optimists will note the remarkable progress that has been made in European research collaboration over the past 50 years and, in particular, over the past 15. Huge EU research programmes have forged active collaboration involving tens of thousands of scientists. Academic mobility between nation states is visibly increasing, everywhere you look.

Almost all major facilities are now planned on the basis of pan-European collaboration. Earlier this month, ground was broken on the latest of these: the European Spallation Source near Lund in Sweden, paid for by 17 European nations. (It is worth noting that the United States has not managed to start work on a billion-dollar-scale research facility for more than a decade.)

Most importantly, a cohesiveness and mutual understanding has emerged between senior European scientists that most parts of the world can only look upon with envy. Compared with the situation in east Asia in particular, the level of everyday dialogue and collaboration that exists in Europe in several major disciplines, such as particle physics and molecular biology, is singularly impressive.

This process had been going on for decades, before the formal concept of the ERA was endorsed by EU heads of state at a summit meeting in Lisbon in 2000.

The idea of taking specific administrative steps to improve researcher mobility was mainly theoretical at first, but has steadily gained impetus. And the decision was taken in 2012 for the European Commission to report annually on ERA progress, with the aim of cajoling more action out of member states.

At the same time, the political context for the ERA initiative has changed for the worse. The ERA was conceived when the EU had just experienced a period of rapid convergence — in particular, economic convergence between living standards in the north and south.

Since 2008, however, the health of Europe's national economies has been diverging. Today, in research and innovation, as in other spheres, the wealthier regions are moving rapidly ahead, with the poorer ones falling behind. On the face of it, this makes the ERA's objectives more elusive than ever.

"With the rapid growth of the Internet, genuine global research collaboration has become almost routine."

Perhaps with this in mind, the commission's latest progress report pulls some of its punches. Earlier talk of 'naming and shaming' those member states that are slowest to implement ERA actions has been reined in.

These actions include steps to improve the portability of researchers' pensions and to address the gender gap in research. Women now obtain around half of Europe's PhDs, but will receive less than a quarter of this year's grants from the prestigious European Research Council. This is a major problem that both universities and research agencies prefer to overlook; its vigorous pursuit is a worthwhile goal for the commission.

Another change that has intruded on the ERA since 2000 is the accelerated emergence of a de facto global research area among elite researchers in most disciplines. Since 2000, with the rapid growth of the Internet, genuine global research collaboration has become almost routine, rendering 'local' collaboration less significant.

Still, Geoghegan-Quinn's successor as research commissioner — the current nominee is Portugal's Carlos Moedas — should pursue the goals of the ERA with as much vigour as possible. There will doubtless be renewed debate in the new European Parliament about the need for a fresh EU directive to force member states' hands over the ERA. In the meantime, it is up to the member states and their institutions to do more.

The 2014 deadline may be about to pass, but the project must endure. Ultimately, its fate rests in the hands of every department, institution and research agency in Europe — to build the ERA, one step at a time. ■

Amped-up plants

Bacterial enzyme supercharges photosynthesis, promising increased yields for crops.

The catalytic conversion of carbon dioxide and water to sugar and oxygen is arguably the most important chemical reaction in the world, and one of the oldest. It is so old, in fact, that it evolved when the world's atmosphere was much lower in oxygen than it is today. So, in a way, photosynthesis is its own worst enemy. Thousands of millions of years later, most modern plants struggle to photosynthesize because of all the darned oxygen in the air — oxygen that they helped to put there. These plants simply cannot distinguish between molecules of carbon dioxide and molecules of oxygen, so they waste their time and energy grabbing both.

Some plants can do better — for example, plenty of weeds (ever wondered why they grow so fast?) have evolved ways to concentrate carbon dioxide inside their leaves, to supercharge their photosynthesis. Cyanobacteria can do this too. But the majority of plants, including most of the crops we rely on for food, have developed a blunter strategy: produce lots and lots of the enzyme that drives the reaction. That enzyme, Rubisco, is thus among the most abundant proteins on the planet.

A significant amount of Rubisco still wastes its time grabbing useless oxygen — reducing the overall efficiency of global photosynthesis by almost one-third. When they discuss ways to boost the world's food supplies, plenty of plant scientists see leaves' wasted photosynthesis capacity as, well, low-hanging fruit.

What if crops could borrow the faster-acting Rubisco system of weeds and cyanobacteria? In theory, this would dramatically boost their growth rate and so their yield, all without needing any extra farmland. The appeal of such a strategy is obvious, particularly in the face of the often-quoted United Nations demand for global food production to double by 2050.

In practice, replacing the enzyme has proved difficult. But there is encouraging news: on *Nature*'s website, researchers report that they have made tobacco plants that use the Rubisco from a cyanobacterium (M. T. Lin *et al. Nature* http://dx.doi.org/10.1038/nature13776; 2014). Sure enough, the transformed plants photosynthesize faster and have higher rates of CO_2 turnover than their conventional counterparts. Faster-growing tobacco plants might not sound like a boon for global welfare, but they do demonstrate what might be possible in future. (Tobacco is a common model organism for genetic-engineering research.)

As biologists Dean Price and Susan Howitt write in an accompanying News & Views (G. D. Price and S. M. Howitt *Nature* http://dx.doi.org/10.1038/nature13749; 2014): "The work is a milestone on the road to boosting plant efficiency. The advance can be likened to having a new engine block in place in a high-performance car engine — now we just need the turbocharger fitted and tuned." Available in any colour you like, as long as it's green. ■

NATURE.COM
To comment online, click on Editorials at:
go.nature.com/xhunqv


CSIRO

UN sustainability goals need quantified targets

Scientists must step up and secure meaningful objectives if they are to protect both people and planet, says **Mark Stafford-Smith.**

The United Nations Millennium Development Goals (MDGs) pass their deadline next year and will be replaced by the broader and more ambitious Sustainable Development Goals (SDGs) to guide world development until 2030.

The SDGs matter because they will set development priorities for governments and businesses, among others. Moreover, they can help to reshape attitudes towards the relationship between economic growth and environmental protection, to help preserve and protect both.

Draft goals were presented to the UN General Assembly last week in New York. A year of negotiations follows, with the final version of the goals scheduled to be affirmed in September 2015. That the world is close to agreeing on a consolidated set of objectives for global sustainability is a game-changer.

However, it is crucial that the new goals are based on the best scientific evidence of environmental problems and the best strategies to mitigate these risks. Scientists have helped to draft the proposed goals, but their input has been weak, fragmented and intermittent. We have less than 12 months to change that.

The first problem is that there are too many proposals: 17 goals encompassing 169 individual targets, ranging from improving maternal health to safeguarding the oceans. The strategy has shifted from a list of priorities to an unwieldy and impractical catch-all. The strength of the original MDGs was their focus.

We should aim for no more than ten goals, with around five or six targets for each. This should offer the right balance between covering enough ground and providing sharp focus. These ten goals should cover social, economic and environmental priorities, and on these points the draft proposals make a good start. Four draft goals discuss global environmental constraints, for climate, water, ecosystems and the oceans. This is a step forward that should be applauded.

Although many of the proposed social targets are ambitious, aspirational and reasonably well defined, the biophysical targets are vague, modest and lack detailed quantification. For example, under the health goal, the first target is specific: "By 2030 reduce the global maternal mortality ratio to less than 70 per 100,000 live births." By contrast, the sustainability target under the food-security goal starts: "By 2030 ensure sustainable food production systems". The target is nebulous and, crucially, omits mention of important constraints on the nitrogen, phosphorus and water cycles. A water target is equally vague: "By 2030, substantially increase water-use efficiency across all sectors".

Such non-specific targets will not provide the integrated framework for people and planet that is so direly needed to drive transformations in energy, resource and land-use systems. Without quantified targets and monitoring, it is impossible to determine whether sufficient progress is being made.

We already know enough about the biophysical systems involved to set specific targets, such as keeping the flow of phosphorus into the ocean to below 11 million tonnes per year.

Perhaps most importantly, the goals must work towards a common purpose. At present, individual goals on energy access and tackling climate change could contradict each other — massive expansion of fossil-fuel use, for example, would satisfy one goal but undermine the other. To prevent this, the goals must be integrated. There are perceived trade-offs between securing the long-term stability and health of the Earth system, and securing water, food and energy security in the short term. But this need not be the case. An integrated approach to food security could also ensure that sustainability targets for nutrient and water cycles are met. For example, we should aim, by 2030, to use no more than 1,000 cubic metres of water per tonne of key food crops produced.

In a similar way, the current potential conflicts between the goals of delivering energy for all and limiting greenhouse-gas emissions can be mediated by strong integrative targets: decrease carbon intensity by increasing the share of renewable energy to 30%, and increase energy intensity by 2.4% per year. Current targets do address these two issues, but without quantification.

These are realistic and achievable changes. But the research community must convince policymakers that such changes are important. Organizations such as the Future Earth initiative, the UN's Sustainable Development Solutions Network and the UN Secretary-General's Scientific Advisory Board must ensure that the right expertise is brought to bear on this challenge at international and regional levels.

At a national level, funding agencies and scientific academies need to bring together expertise to support this international process. Scientists should identify and talk to the negotiators who will finalize the draft goals.

2015 is a significant year for international politics related to global change. Nations will also agree on a new climate deal and a strategy for disaster-risk reduction. Traditionally, science has struggled to respond flexibly to the demands and speed of some political processes. But the SDGs are too important for the research community to let the opportunity pass. ■

THE STRATEGY HAS SHIFTED FROM A LIST OF PRIORITIES TO AN UNWIELDY AND IMPRACTICAL CATCH-ALL.

➲ **NATURE.COM**
Discuss this article online at:
go.nature.com/g5rzf1

Mark Stafford-Smith *is chair of the Science Committee of Future Earth, and principal research scientist with the Commonwealth Scientific and Industrial Research Organisation in Canberra, Australia.*
e-mail: mark.staffordsmith@csiro.au

RESEARCH HIGHLIGHTS

Selections from the scientific literature

VIROLOGY

Wide area of Ebola risk in Africa

The region in Africa at risk of an outbreak of the Ebola virus is larger than previously thought.

Simon Hay at the University of Oxford, UK, and his team mapped data from 23 Ebola outbreaks in humans, including the current one, and 51 reports of Ebola in other animals. They combined the data with information on human mobility and the range of animal hosts suspected of carrying the virus, such as Old World fruit bats.

The team found that the potential reservoir for the virus spans 22 countries in western and central Africa, and includes an area containing more than 15 million people, where Ebola cases have already occurred.

This finding, combined with other recent trends such as increasing urbanization, may account for an apparent increase in the frequency and size of outbreaks since 2000.

eLife **http://doi.org/vms (2014)**

DANIEL KARP

ECOLOGY

Bird diversity at risk from farming

Birds that have the longest evolutionary history are also the most threatened by agriculture.

Luke Frishkoff at Stanford University in California, Daniel Karp at the University of California, Berkeley, and their team studied 12 years of bird survey data, covering nearly 500 species from three types of land use in Costa Rica: forests, diversified agriculture and intensive farming of just a few crop species. They found that on farmland, evolutionarily distinct birds, which are related to few other living species — such as the rufous-tailed jacamar (*Galbula ruficauda*; pictured) — went extinct locally at higher rates than those that had evolved more recently.

However, less-intensive agriculture fostered greater levels of phylogenetic diversity than intensive farming, so the authors suggest that this type of agriculture could help to conserve some bird evolutionary history.

Science 345, **1343–1346 (2014)**

PALAEONTOLOGY

How a shark used its saw-like jaw

Despite having a set of teeth shaped like a circular saw, an extinct shark probably devoured only soft-bodied prey.

The spiral-shaped tooth arrangement (**pictured**) of *Helicoprion davisii*, an animal some 300 million years old, has puzzled palaeontologists for more than a century. Last year, a team determined that the teeth were surrounded by the shark's lower jaw.

In a follow-up study, Jason Ramsay at the University of Rhode Island in Kingston and his team used computed tomography scans of the fossils to reconstruct the jaw muscles and model the mechanics of the jaw and teeth to determine how and what the animal ate. They concluded that older teeth at the front of the jaw snagged prey whereas younger, stronger teeth deeper in finished them off.

The shark teeth were rarely worn or broken, suggesting the animals ate soft-bodied sea creatures such as cephalopods.

J. Morphol. **http://dx.doi.org/10.1002/jmore.20319 (2014)**

GLACIOLOGY

Surface heat led to ice-shelf demise

The collapse of Antarctica's giant Larsen B Ice Shelf in 2002 was probably caused by warming at the surface rather than by instability at the bottom of the ice sheet.

Eugene Domack at the University of South Florida in St Petersburg and his colleagues mapped the sea floor below where the shelf used to be. They also analysed marine sediment cores to reconstruct characteristics of the ice shelf's grounding zone — where the floating ice shelf meets underlying bedrock — before the ice collapsed.

They found that this zone had remained stationary for some 12,000 years, challenging the idea that structural changes at the bottom of the ice shelf might have caused Larsen B's

IMNH

disintegration.

The findings could inform estimates of how much Antarctic melting will contribute to future sea level rise, the authors say.
Science 345, **1354–1358 (2014)**

IMMUNOLOGY

The gut improves vaccine effects

Bacterial residents of the gut boost immune responses to vaccination in mice.

Humans vaccinated against the influenza virus ramp up expression of a protein called TLR5, which is involved in detecting certain types of bacterium. To see how this protein and gut bacteria might affect immune responses to vaccines, Bali Pulendran of Emory University in Atlanta, Georgia, and his colleagues studied mice that lack the gene encoding TLR5.

They found that the animals produced fewer antibodies in response to flu vaccination than normal mice. The team saw similar effects in mice reared in a germ-free environment and in those treated with powerful antibiotics. Antibody responses could be restored, however, by inoculating the mice with the kind of bacteria to which TLR5 is sensitive.

The results suggest that antibiotic treatment could hinder the effects of certain vaccines, the authors say.
Immunity **http://doi.org/vm3 (2014)**

PARTICLE PHYSICS

Better estimate of Higgs mass

Researchers have decreased the uncertainty of their estimate of the mass of the Higgs boson, the particle thought to bestow mass to matter.

The ATLAS collaboration, one of two teams that detected the Higgs at the Large Hadron Collider near Geneva, Switzerland, reanalysed data and improved detector calibration to come up with the revised mass of 125.36 gigaelectronvolts (GeV), with a systematic uncertainty of 0.18 GeV — an improvement by a factor of three.

The measurement will refine predictions of the Higgs' behaviour and help to identify potential phenomena not predicted by the standard model of physics, the team says.
Phys. Rev. D 90, **052004 (2014)**

INFECTIOUS DISEASE

Mosquitoes awaken malaria

Mosquitoes biting a malaria-carrying host coax the malaria parasite to come out of hiding, resulting in greater disease transmission.

Sylvain Gandon at the National Centre of Scientific Research in Montpellier, France, and his colleagues infected canaries (*Serinus canaria*) with a malaria parasite that is specific to birds (*Plasmodium relictum*), and then exposed them to mosquitoes that were not carrying the parasite.

After the birds were bitten by malaria-free insects, the level of parasites rose in the birds' blood. Mosquitoes that subsequently bit birds were more likely to pick up and transmit the parasite than insects attacking birds that had not been initially bitten.

The researchers conclude that mosquito bites trigger *Plasmodium* to emerge from its dormant stage.
PLoS Pathog. 10, **e1004308 (2014)**

MICROBIOLOGY

Vaginal microbe makes drug

A bacterium that lives in the human vagina produces an antibiotic, suggesting how the microbiome could be mined for possible drug candidates.

Michael Fischbach at the University of California, San Francisco, and his colleagues trained a computer program to recognize genes that are known to make molecules that could be used as drugs, and then asked the program to hunt for similar genes in the human microbiome.

This yielded thousands of genes, including some that make a class of antibiotics called thiopeptides. The team isolated a new thiopeptide from a vaginal microbe grown in the lab, and found that the compound could kill the same types of bacterium as other thiopeptides.

This could be the first drug discovered in and isolated from an organism living in humans, the authors say.
Cell 158, **1402–1414 (2014)**

ECOLOGY

Sneaky ants steal in plain sight

A recently discovered parasitic ant species steals food from colonies of another ant by disguising itself as the host.

Scott Powell at George Washington University in Washington DC and his co-workers discovered the parasitic ant, *Cephalotes specularis* (**pictured** right), in the Brazilian woodland savannah. *C. specularis* lives only with its host, the highly aggressive *Crematogaster ampla* ant (**pictured** left).

SCOTT POWELL

The researchers found that, rather than introduce its brood into the host's nest like other parasitic ants, *C. specularis* mimics the body posture of the host worker ants to move freely around the host's territory. The deceptive ant follows the host's pheromone trails to locate food, and manages to sneak undetected into 89% of potential host territories.
Am. Nat. **http://dx.doi.org/10.1086/677927 (2014)**

SOCIAL SELECTION

Popular articles on social media

High retraction rates raise eyebrows

Amid a wave of recent retractions, researchers are taking to social media to discuss a perennial favourite: a three-year-old paper looking at the relationship between a journal's impact factor and its retraction frequency. The 2011 report proposed a "retraction index", a measure of the likelihood that a paper in a given journal will eventually be pulled from the literature. The authors looked at articles published from 2001 to 2010 in 17 journals and plotted the journals' retraction indexes against their impact factor. The result was clear: the higher the impact factor, the higher the retraction index. "You know 'high impact' journals? All that means is that work is more likely to be retracted," tweeted Jon Tennant, who studies palaeontology at Imperial College London, earlier this month. David Basanta, a cancer researcher at the Moffitt Cancer Center in Tampa, Florida, responded on Twitter: "A case could be made that more people try to replicate the results."
Infect. Immun. 79, **3855–3859 (2011)**

Based on data from altmetric.com. Altmetric is supported by Macmillan Science and Education, which owns Nature Publishing Group.

NATURE.COM
For more on popular papers:
go.nature.com/bfqx3j

NATURE.COM
For the latest research published by *Nature* visit:
www.nature.com/latestresearch

SEVEN DAYS *The news in brief*

FUNDING

Science partnership

The United Kingdom and South Africa signed a multi-year science and technology partnership on 9 September, to be jointly funded with £7.8 million (US$12.7 million) annually. Research priorities include public health, food security and technology development. The money will be administered by the Newton Fund, established by Britain to support science collaborations with developing countries (see go.nature.com/yhmmvp). The two countries also announced three-year partnerships to study tuberculosis and non-communicable diseases in Africa.

FACILITIES

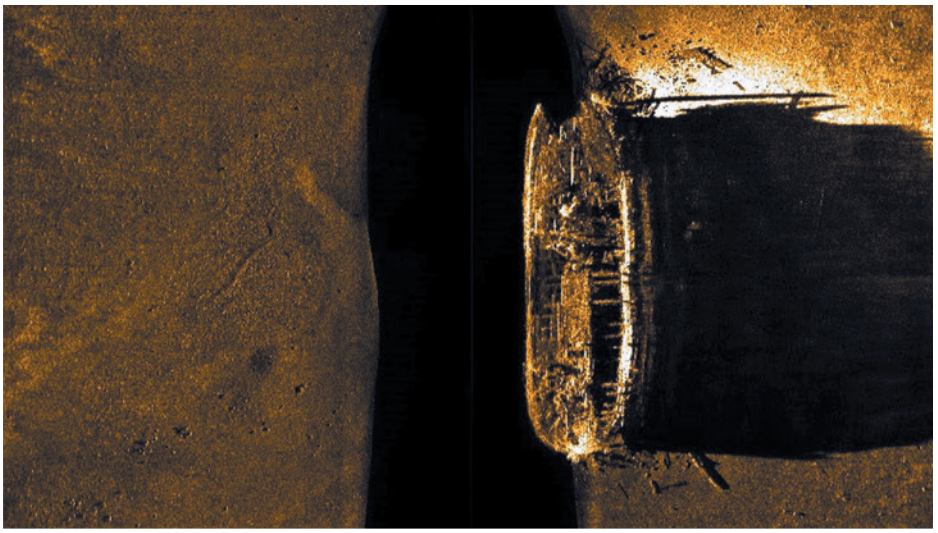
University blaze

A fire at the University of Nottingham, UK, on 12 September destroyed a carbon-neutral chemistry laboratory that was under construction. Funded in part by a £12-million (US$19-million) grant from drug giant GlaxoSmithKline, the facility was slated to open next year. The university says that it is working with GlaxoSmithKline and its contractor to develop a plan to rebuild the laboratory.

RESEARCH

Data falsified

The BMJ Publishing Group announced last week the retraction of a June 2013 article in the *British Journal of Psychiatry* that reported a higher incidence of epigenetic changes in people with bipolar disorder who had experienced early-life trauma. An investigation by the University of Geneva in Switzerland found that senior author Alain Malafosse had fabricated the DNA-methylation data underlying the paper's conclusions. A former director of genetic psychiatry in the university's hospital system, Malafosse is also accused of embezzling 1.7 million Swiss francs (US$1.8 million) in government research funds.



PARKS CANADA

Long-lost ship found in Canadian Arctic

Archaeologists have found one of the ships from the Franklin expedition — which disappeared in the 1840s — off King William Island in the Canadian Arctic. A team led by Parks Canada last week discovered either HMS *Erebus* or HMS *Terror*, Canadian Prime Minister Stephen Harper announced on 9 September. A remotely operated vehicle helped to locate the ship's remains from a sonar image (pictured). Since 2008, Parks Canada has scoured hundreds of square kilometres of ocean floor in search of the ships, which British explorer John Franklin commanded while seeking the Northwest Passage. Historical records suggest that some of the explorers died while the ships were trapped in ice, and others perished while attempting to walk south. See go.nature.com/ugcvuy for more.

Curiosity arrives

After more than 2 years and 9.5 kilometres of driving across Mars, NASA's Curiosity rover has reached its ultimate goal: a peak called Mount Sharp. NASA announced the milestone on 11 September following the vehicle's arrival at a smoother, less-cratered rock formation known to surround the mountain. The agency also defended the mission against recent criticism from a senior review panel, which suggested that the project lacks clearly defined science goals and spends too little time collecting and analysing data.

Bad news for birds

Many US bird populations are declining, largely owing to disruption of habitats by human development, according to *The State of the Birds 2014*, a long-term study published on 9 September by organizations including the Smithsonian Institution in Washington DC and the US Geological Survey in Reston, Virginia. Birds in the arid habitats of the western United States have been hardest hit, with some areas seeing a 46% population loss since 1968. The report names 230 at-risk species, including the Laysan albatross (*Phoebastria immutabilis*), a seabird whose low-elevation breeding areas are threatened by rising sea levels.

Brain-project talks

The European Union's flagship €1-billion (US$1.6 billion) Human Brain Project has begun a 'mediation process' after hundreds of neuroscientists fiercely criticized the project's management and scientific direction in July (see *Nature* **511,** 133–134; 2014). Germany's national Jülich Research Centre said on 12 September that its board chairman Wolfgang

Marquardt will lead the mediation, which should be completed by mid-2015.

Stem-cell test

On 12 September, a woman in Japan became the first person to receive an experimental treatment derived from induced pluripotent stem cells — cells reprogrammed from mature tissue to be capable of becoming many types of cell. See page 287 for more.

POLICY

Reef protection

Australia announced on 15 September how it intends to safeguard the troubled Great Barrier Reef, which faces threats from climate change, coral-eating starfish and industrial development. The proposed plan outlines actions that the national and regional governments must take over the next 35 years, and includes ways to improve water quality and biodiversity. Conservationists have already said that the plan does not offer enough protection; the proposal is under consultation until 27 October.

Shark cull called off

Western Australia has halted the extension of a controversial shark cull, after environmental regulators warned in a report on 11 September of scientific uncertainty surrounding the programme's effects on the great white shark — a protected species. In response to seven fatal shark attacks on beach-goers between 2010 and 2013, baited traps called drum lines were set up earlier this year, catching more than 172 sharks — most of them tiger sharks (**pictured**). State officials had sought to extend the programme for another three years.

SEA SHEPHERD/ANIMAL AMNESTY

Swiss diplomacy

Scientists in Switzerland are once more being allowed to compete for funds from the European Union's Horizon 2020 research programme. Switzerland was excluded in February after it imposed curbs on the immigration of citizens of the European Union (see *Nature* **506,** 277; 2014). The European Commission announced on 12 September that it has negotiated temporary exceptions to allow applications for some grants up to the end of 2016. That includes European Research Council grants, Marie Skłodowska-Curie fellowships and specific large projects, such as the Human Brain Project (headquartered in Lausanne) and nuclear-fusion programmes.

Nuclear restart

Japan has taken its first steps towards restarting nuclear power generation, following the Fukushima Dai-ichi plant meltdown in 2011. On 10 September, the Nuclear Regulation Authority granted safety approval to the Sendai Nuclear Power Station, concluding that the facility's two reactors satisfied new regulations designed to guard against disasters such as earthquakes and tsunamis. The plant must still clear further safety checks and obtain local-government approval before being turned on. Japan's 48 other reactors remain offline.

PEOPLE

EU science leaders

European Commission president-elect Jean-Claude Junker announced on 10 September his nominations for the 28 members of the next commission. The choices include former engineer and economist Carlos Moedas from Portugal as commissioner for research, science and innovation; and Miguel Arias Cañete, Spain's former agriculture and environment minister, for climate and energy commissioner. The nominations must now be approved by the European Parliament. See go.nature.com/t73tfi for more.

COMING UP

21–24 SEPTEMBER
Two spacecraft are due to enter orbit around Mars: first NASA's MAVEN craft and then, three days later, India's Mars Orbiter Mission. See page 291 for more.

23 SEPTEMBER
A major United Nations Climate Summit kicks off in New York City, intended to stoke enthusiasm for future international negotiations. See page 289 for more.

BUSINESS

Stem-cell stock

On 10 September, the US Securities and Exchange Commission (SEC) charged a prominent stem-cell company and a former employee of the firm with defrauding investors. The SEC says that Gary Rabin, former chief executive of Advanced Cell Technology (ACT) in Marlborough, Massachusetts, waited too long to notify investors that he had sold US$1.5 million of the firm's stock between 2010 and 2012. The company is struggling to raise funds to support its ongoing research (see *Nature* http://doi.org/q8f; 2014). ACT and Rabin agreed to settle the charges without admitting or denying them.

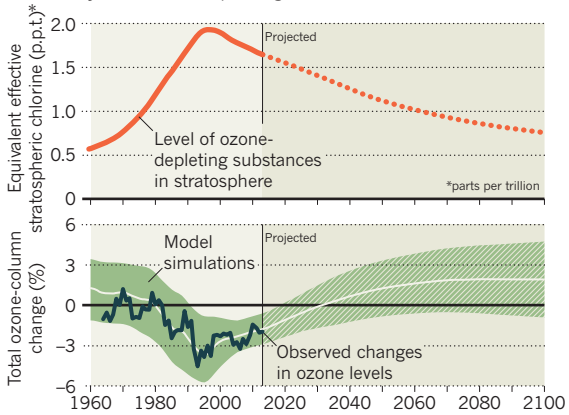
TREND WATCH

The ozone layer seems to be on the mend, says a 10 September report by the United Nations Environment Programme and the World Meteorological Organization. Atmospheric levels of ozone-depleting chemicals, including chlorofluorocarbons — once widely used in refrigerants, and restricted under the 1987 Montreal Protocol — have declined by 10–15% in the past 10–15 years. In parts of the atmosphere, ozone levels have increased by 5% since 2000. See go.nature.com/ozr4np for more.

SOURCE: UNEP/WMO

OZONE ON THE REBOUND

Atmospheric levels of chemicals that destroy ozone have declined by 10–15% since peaking around 2000.



NATURE.COM
For daily news updates see:
www.nature.com/news

NEWS IN FOCUS

JIJI PRESS/AFP/GETTY

Masayo Takahashi is the first to implant tissue derived from induced pluripotent stem cells into a person.

REGENERATIVE MEDICINE

Japan stem-cell trial stirs envy

Researchers elsewhere can't wait to test iPS cells in humans.

BY SARA REARDON & DAVID CYRANOSKI

"It's awesome, it's amazing, I'm thrilled, I've been waiting for this," says Jeanne Loring, a stem-cell biologist at the Scripps Research Institute in La Jolla, California. She is one of several researchers around the world to welcome the news that a Japanese woman with visual impairment had become the first person to receive a therapy derived from stem cells known as induced pluripotent stem (iPS) cells.

A lot rides on this trial. If the procedure proves safe, it could soften the stance of regulatory bodies in other nations towards human trials of iPS cells, and it could pave the way for treatments for other conditions, such as Parkinson's disease and diabetes. It could also cement Japan, recently plagued by a stem-cell scandal, as a frontrunner in iPS-cell research.

Pioneered in 2006 by Shinya Yamanaka, now director of the Center for iPS Cell Research and Applications at Kyoto University, iPS cells are created by inserting certain genes into the DNA of adult cells to reprogram the cells back to an embryonic-like state. The cells can then be turned into almost any tissue type, much as embryonic stem cells can. But because iPS cells can be derived from a patient's own tissue, the hope is that they will dodge some of the controversial aspects and safety concerns of those derived from embryos.

In 2012, Yamanaka received a Nobel prize for his work, and the field has now matured, with teams across the world champing at the bit to test therapies based on iPS cells in people. Loring, for example, uses the cells to create dopamine-producing neurons as a potential therapy for Parkinson's disease, and says that she will start clinical trials as soon as the US Food and Drug Administration (FDA) gives the go-ahead.

Still, tissues made from iPS cells carry their own concerns, and that had stopped any country from approving them for a clinical trial. The body's immune system could attack them, or they might contain some cells that are still in the pluripotent state and cause cancerous growths — although Loring points out that this has not happened with human trials of therapies based on embryonic stem cells, for which the same concerns would apply.

A GREEN LIGHT

In July 2013, however, Japan's regulatory authorities gave the go-ahead for a team led by ophthalmologist Masayo Takahashi at the RIKEN Center for Developmental Biology (CDB) in Kobe to collect cells to be used in a clinical iPS-cell pilot study.

Her team took skin cells from the first patient, a woman in her seventies who had retinal damage owing to a condition known as age-related macular degeneration. The researchers then reprogrammed the skin cells into iPS cells and coaxed the unspecialized cells into becoming retinal tissue. On 8 September, Takahashi provided evidence that those cells were genetically stable and safe, a prerequisite for them to be transplanted into the eye. The procedure took place four days later, and RIKEN has reported that the patient experienced no serious side effects.

In this instance, the woman's vision is unlikely to improve. However, researchers around the world are watching to see whether the cells stop the retina from deteriorating further and whether any side effects develop. Should the woman experience serious consequences, iPS-cell research could be set back years, much as gene therapy was in 1999 when a patient died in a trial that attempted to use ▶

a modified gene to correct a type of liver disease. "That wakes me up at night," Loring admits.

If Takahashi's trial succeeds, however, it could send a powerful signal to other regulatory agencies such as the FDA and the European Medicines Agency. "If Masayo can demonstrate that these cells are safe in patients, that will have calmed some of the anxiety about the new cell type out there," says developmental molecular biologist Kapil Bharti at the National Eye Institute in Bethesda, Maryland. Bharti is leading an effort within the US National Institutes of Health (NIH) to develop an iPS-cell therapy for macular degeneration using an approach similar to Takahashi's. He hopes to apply to the FDA in 2017 to begin clinical trials.

"They are sort of envious because you can move forward rapidly in Japan."

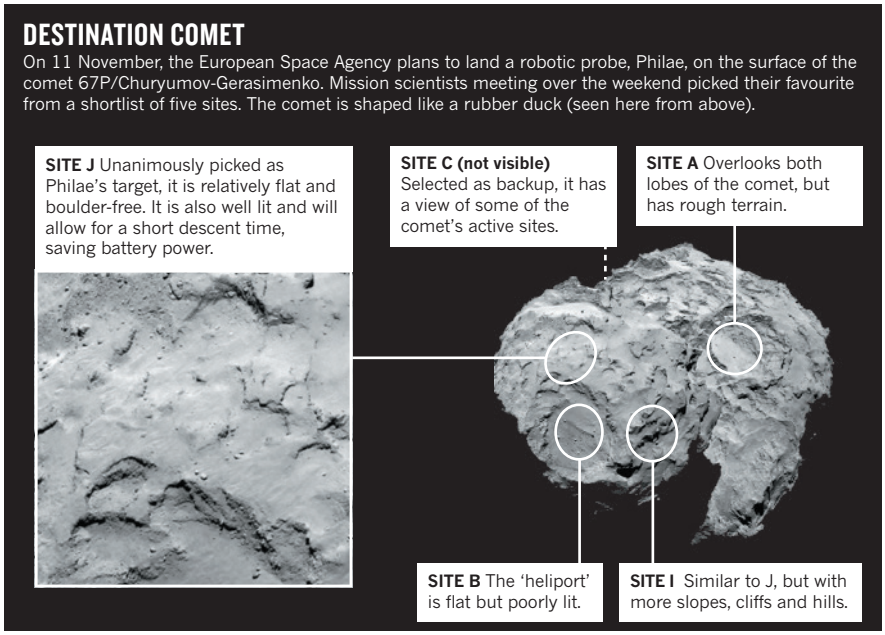
Others are less patient. Stem-cell biologist Mahendra Rao, who until recently headed the NIH Center for Regenerative Medicine that backs Bharti's trial and is now at the New York Stem Cell Foundation, says that regulations have been moving too slowly for companies outside Japan that want to do similar trials. One of these is Q Therapeutics in Salt Lake City, Utah, which he founded and which is developing cell-based therapies for neurodegenerative diseases. "They are sort of envious because you can move forward rapidly in Japan," he says.

But the Japanese system is also controversial. Since approving the Takahashi study, regulators, keen to stay ahead in stem-cell research, have changed the law to make it easier to test therapies based on iPS cells clinically, a move that some say could result in ineffective treatments being thrust on desperate patients.

The surgery offers some welcome positive news for RIKEN, and Japan, in the wake of a stem-cell scandal and tragedy. "It gives them some of the credibility back," Loring says.

Earlier this year, researchers from RIKEN's CDB published two papers in *Nature* claiming to have made stem cells through a technique known as stimulus-triggered acquisition of pluripotency, or STAP. The papers were retracted in July, leading to a misconduct charge for one researcher, and contributing to the suicide of another. The CDB is now being halved in size, but in 2016, RIKEN is planning to open the ¥3-billion (US$28-million) Kobe Eye Center to develop cutting-edge procedures such as Takahashi's. And the nation's new legislation means that several other Japanese researchers are expected to begin clinical iPS-cell studies soon, including Takahashi's husband, Kyoto University's Jun Takahashi, who is planning a trial for Parkinson's disease.

Outside Japan, many researchers hope that Masayo Takahashi's trial will hasten the translation of their own work to therapies. "Every time someone goes down this path, it is easier for those who are following," Loring says. ■



ESA/ROSETTA/MPS FOR OSIRIS TEAM MPS/UPD/LAM/IAA/SSO/INTA/UPM/DASP/IDA

SPACE

Lander to aim for comet's 'head'

Touchdown site for Rosetta probe chosen unanimously.

BY ELIZABETH GIBNEY

There is no easy way to alight on a 4-kilometre-long, rubber-duck-shaped ice ball that is spinning as it flies through the outer Solar System. But scientists working on the European Space Agency's Rosetta mission have selected a spot on the 'head' of the comet, called 67P/Churyumov-Gerasimenko, that they think will give them the best chance for gently landing Philae, a washing-machine-sized robotic probe.

Planned for 11 November, the first soft landing ever attempted on a comet is fraught with risk. When researchers still thought that the object had a regular, potato-like shape, they estimated the landing's chance of success at 70–75%. Now that the Rosetta orbiter has taken a closer look and revealed the curious shape, the odds are lower. Mark McCaughrean, a senior science adviser at the European Space Agency (ESA) directorate of science and robotic exploration in Noordwijk, the Netherlands, puts them at roughly "fifty–fifty".

The mission scientists were unanimous in their choice of landing spot — a 1-square-kilometre patch known as site J — from a shortlist of five (see 'Destination comet'). Philae lead scientist Jean-Pierre Bibring of the University of Paris-South in Orsay says that site J emerged as the favourite after the first day of a meeting held on the weekend of 13–14 September at the French National Centre for Space Studies in Toulouse. "This site is not the best for every one of the technical and scientific criteria, but overall it's by far the best for mission success," he says.

In a precisely choreographed fly-by, Rosetta will release Philae from a distance of about 10 kilometres. From there, the probe will drift unguided towards the target, where it will secure itself with harpoons and screws and start work. The information that Philae collects about the comet's innards will help to calibrate data gathered by the more powerful instruments on Rosetta, says McCaughrean. "There are many things that we can only do on the surface," he says.

A major advantage of site J is that the drop from Rosetta will be relatively short, at just 7 hours. That means that Philae will have more battery power to run its instruments after landing, because it will take two days to recharge using its solar panels.

The region also has relatively few boulders that could capsize Philae on landing. Still,

nowhere is devoid of danger. "There is no one big Heathrow airport on the surface where you can say, 'No problem'," says McCaughrean.

Although chosen mainly for technical considerations, the site is also interesting scientifically. It is just a few hundred metres from two pits that scientists think will become more active, spewing out gas and dust, as the comet moves closer to the Sun and heats up. The position of the landing relative to Rosetta's orbit will also afford the best chance of transmitting radio waves between the two craft to map the comet's interior, says Bibring.

The mission team says that it reached its decision quickly, then spent most of the meeting's second day picking a backup — a spot on the comet's body known as site C.

Other potential backups included a crater nicknamed 'the heliport' for its flatness, but the site is not as well lit as site C. And a spot that would have provided views of the body, head and highly active 'neck' region had been effectively ruled out before the meeting even started, says Bibring, because Rosetta would have needed to drop to an orbit that was dangerously close to the comet.

The Rosetta team is rushing to gather as much data as possible and to stick to the November landing date, because after that increased comet activity could damage the orbiter.

Rosetta has been chasing its quarry for a decade. After waking from hibernation in January, it arrived at its destination in August and has been charting its target from ever-shrinking orbits ever since. Rosetta will continue to follow the comet as it journeys around the Sun. ■

POLICY

Climate summit previews push for new global treaty

United Nations meeting aims to spark enthusiasm for a 2015 emissions pact.

BY LAUREN MORELLO

When the United Nations Climate Summit begins on 23 September in New York City, US President Barack Obama will be there. But many of his counterparts from other major greenhouse-gas-emitting countries — including China, India, Germany and Australia — plan to stay at home. Still, the meeting could offer important clues to how a UN push to forge a new international climate pact by the end of 2015 will play out.

Approved in 1992 at the Earth Summit in Rio de Janeiro, the Kyoto Protocol is the only legally binding international treaty governing greenhouse-gas emissions. Parts of it expired at the end of 2012, and efforts have since been afoot to develop a new pact, as demanded by the 22-year-old United Nations Framework Convention on Climate Change (UNFCCC).

The New York meeting is not part of the formal process to shape a new international climate treaty. Instead, the gathering was conceived by UN secretary general Ban Ki-moon as a way of marshalling enthusiasm for the effort, which is set to conclude in Paris in December 2015; any agreement would take effect in 2020.

International climate negotiations have a long and chequered history. The Kyoto Protocol was never ratified by the United States, did not require rapidly developing countries to reduce their emissions and was rejected by Canada and Russia as the first round of emissions-reductions commitments expired in 2012.

In the meantime, the world's output of carbon dioxide and other heat-trapping gases has continued to rise. The level of CO_2 in the atmosphere reached 396 parts per million in 2013, 42% higher than pre-industrial levels. Last year's was the largest annual increase since 1984, according to figures reported on 9 September by the World Meteorological Organization in Geneva, Switzerland.

Climate negotiators face sticky questions as they work to craft a new agreement. Some are basic: will the Paris treaty be legally binding? Others are more complex and long-standing. Many developing nations, for instance, worry that carbon cuts will jeopardize their economic progress. The challenge is formidable, says Nicholas Stern, a climate-change economist at the London School of Economics. By 2030, Stern says, the world must reduce its greenhouse-gas emissions by roughly 20% from the current level to have a chance of limiting warming to 2 °C above pre-industrial temperatures, the UNFCCC's stated goal. Current emissions pledges put the world on track for a 3 °C warming by 2100, according to a 7 September report by PriceWaterhouseCoopers.

Yet climate-policy experts insist that there is reason for optimism heading into the New York meeting. "We need to stop judging climate action by whether we get a so-called legally binding treaty," says Paul Bledsoe, senior fellow at the German Marshall Fund in Washington DC and a White House climate-change official under former president Bill Clinton.

Bledsoe sees signs of meaningful progress by major emitters such as China. The country has enacted a cap-and-trade programme to reduce greenhouse-gas emissions in seven provinces and increased its investment in renewable energy. The world's leading greenhouse-gas emitter is also considering a ban on coal-fired electricity generation in the Beijing area to address air-quality concerns. The move would be an important step away from the form of power generation that produces the most greenhouse-gas emissions.

There are also signs of progress in the United States, the second-largest emitter of greenhouse gases, says David Waskow, director of the international climate project at the World Resources Institute in Washington DC. Obama is using his executive power to enact policies that reduce greenhouse gases — bypassing Congress, which has long stymied any plan for emissions reductions. Waskow points to Obama's proposal, unveiled in June, to cut greenhouse-gas emissions from existing power plants, which produce 38% of the country's total. The United States already regulates emissions from automobiles and is working to finalize limits for new power plants.

Last year's was the largest annual CO_2 increase since 1984.

Although India's prime minister, Narendra Modi, will not be attending, the stance of India's representatives to the New York summit could yield clues to the country's climate policy. Modi took office in May and championed renewable energy during more than a decade as chief minister of Gujarat state. "It's a new government, and they haven't had a lot of time to sink their teeth into this," says Jake Schmidt, director of the international climate programme at the Natural Resources Defense Council in Washington DC. "India is a big unknown."

But the true test of progress towards a new treaty will not come until March, when nations are required to submit their national greenhouse-gas reductions goals to the UNFCCC. "That is really the timeline to keep an eye on," says Waskow. ■

BEBETO MATTHEWS/AP

Soft drinks are just some of the many products that use artificial sweeteners.

NUTRITION

Sugar substitutes linked to obesity

Artificial sweetener seems to change gut microbiome.

BY ALISON ABBOTT

The artificial sweeteners that are widely seen as a way to combat obesity and diabetes could, in part, be contributing to the global epidemic of these conditions.

Sugar substitutes such as saccharin might aggravate these metabolic disorders by acting on bacteria in the human gut, according to a study published by *Nature* this week (J. Suez *et al. Nature* http://dx.doi.org/10.1038/nature13793; 2014). Smaller studies have previously purported to show an association between the use of artificial sweeteners and the occurrence of metabolic disorders. This is the first work to suggest that sweeteners might be exacerbating metabolic disease, and that this might happen through the gut microbiome, the diverse community of bacteria in the human intestines. "It's counter-intuitive — no one expected it because it never occurred to them to look," says Martin Blaser, a microbiologist at New York University.

The findings could cause a headache for the food industry. According to BCC Research, a market-research company in Wellesley, Massachusetts, the market for artificial sweeteners is booming. And regulatory agencies, which track the safety of food additives, including artificial sweeteners, have not flagged such a link to metabolic disorders. In response to the latest findings, Stephen Pagani, a spokesman for the European Food Safety Authority (EFSA) in Parma, Italy, says that, as with all new data, the agency "will decide in due course whether they should be brought to the attention of panel experts for review".

A team led by Eran Elinav of the Weizmann Institute of Science in Rehovot, Israel, fed mice various sweeteners — saccharin, sucralose and aspartame — and found that after 11 weeks, the animals displayed glucose intolerance, a marker of propensity for metabolic disorders.

To simulate the real-world situation of people with varying risks of these diseases, the team fed some mice a normal diet, and some a high-fat diet, and spiked their water either with glucose alone, or with glucose and one of the sweeteners, saccharin. The mice fed saccharin developed a marked glucose intolerance compared to those fed only glucose. But when the animals were given antibiotics to kill their gut bacteria, glucose intolerance was prevented. And when the researchers transplanted faeces from the glucose-intolerant saccharin-fed mice into the guts of mice bred to have sterile intestines, those mice also became glucose intolerant, indicating that saccharin was causing the microbiome to become unhealthy.

Elinav's team also used data from an ongoing clinical nutrition study that has recruited nearly 400 people in Israel. The researchers noted a correlation between clinical signs of metabolic disorder — such as increasing weight or decreasing efficiency of glucose metabolism — and consumption of artificial sweeteners.

But "this is a bit chicken-and-egg", says Elinav. "If you are putting on weight, you are more likely to turn to diet food. It doesn't necessarily mean the diet food caused you to put on weight."

So his team recruited seven lean and healthy volunteers, who did not normally use artificial sweeteners, for a small prospective study. The recruits consumed the maximum acceptable daily dose of artificial sweeteners for a week. Four became glucose intolerant, and their gut microbiomes shifted towards a balance already known to be associated with susceptibility to metabolic diseases, but the other three seemed to be resistant to saccharin's effects. "This underlines the importance of personalized nutrition — not everyone is the same," says Elinav.

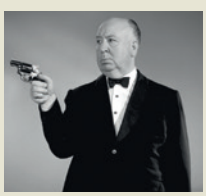
He does not yet propose a mechanism for the effect of artificial sweeteners on the microbiome. But, says Blaser, understanding how these compounds work on some species in the gut might "inspire us in developing new therapeutic approaches to metabolic disease".

Yolanda Sanz, a nutritionist and vice-chair of the EFSA's panel on dietetic products, nutrition and allergies, says that it is too soon to draw firm conclusions. Metabolic disorders have many causes, she points out, and the study is very small. ■

MORE ONLINE

CBS PHOTO ARCHIVE/GETTY

TOP STORY



'Vegetative' mind responds to Hitchcock thriller go.nature.com/f166wd

MORE NEWS

- Q&A: UK medical charities chief Aisling Burnand talks to *Nature* go.nature.com/bj7pek
- Novel antibiotic from vaginal bacterium go.nature.com/9absfg
- Fresh details on retracted stem-cell papers go.nature.com/mc4hrk

NATURE PODCAST

Artificial sweeteners and health, ageing migrating birds, and informed consent in mental health nature.com/nature/podcast

SPACE

Indian Mars craft prepares for orbit

Mangalyaan aims to be Asia's first successful Martian mission.

BY SANJAY KUMAR

VIPUL M. BARAD

The Mars Orbiter Mission will hunt for methane.

Vignesh Nair wanted to know the speed of the spacecraft; Mayyan Baatish asked why India was going to Mars at all, given the cost. For once, the 30-odd members of the Astronomy Club at the G. D. Goenka Public School in a Delhi suburb were discussing something home-grown: India's Mars spacecraft Mangalyaan, which is due to start orbiting the red planet on 24 September. If all goes according to plan, it will be the first successful Mars mission launched by an Asian nation — and a point of Indian pride.

Mangalyaan, known formally as the Mars Orbiter Mission, or MOM, was launched by the Indian Space Research Organisation (ISRO) last November. With 5 scientific instruments that collectively weigh just 15 kilograms, it is designed to image the planet and probe the composition of the surface and atmosphere, including testing for methane and measuring the ratio of deuterium to hydrogen.

Those are modest goals compared with, say, the much larger NASA orbiter MAVEN (Mars Atmosphere and Volatile Evolution), which is also en route to the red planet. Scheduled to arrive just three days ahead of MOM, it has eight instruments and would be the first spacecraft to examine questions such as how the solar wind has stripped away the Martian atmosphere.

As a result, the anticipation surrounding MOM comes not from the science, but from what a safe arrival would mean for India. "It will be a validation that Indian research and development has come of age," says Amitabha Ghosh, an Indian-born planetary scientist based in Washington DC. "India is still perceived as a place where work is outsourced, not because of superior science and engineering skills but because of a cost advantage."

ISRO has launched 35 satellites for countries including France, Germany, Canada, Israel and Singapore. Success for MOM could boost that commercial space industry, says Ajey Lele, a research fellow at the Institute for Defence Studies and Analysis, a think tank in New Delhi.

But trepidation still dogs the mission. MOM is about to enter a critical period: it has been in sleep mode for several months and must soon restart and then slow itself down, by firing its rockets for about 24 minutes, before it can enter Martian orbit. There is only one opportunity for insertion, says A. S. Kiran Kumar, who runs ISRO's Space Applications Centre in Ahmedabad. "We are verifying everything daily and watching closely for any disturbance," he adds.

Ghosh worries that MOM's development was rushed, having taken just 15 months, according to ISRO. "A significant gestation period would have ensured proper engineering rigour and maximized the chances of success," he says.

Scientists also question the choice of rocket for the mission. MOM was launched using ISRO's low-power workhorse rocket, the Polar Satellite Launch Vehicle (PSLV), designed for putting satellites into low Earth orbit. That limited the weight of the payload — and, some suggest, MOM's scientific potential, although ISRO says that it miniaturized components to compensate. ISRO has been working for more than a decade and a half on the more powerful Geosynchronous Launch Vehicle, but the rocket has suffered a string of failures and did not manage to conduct a fully successful launch until January this year.

Whatever happens next, says Lele, India has learnt a lot from MOM. Using the lightweight PSLV at launch, for example, meant that the craft had to take a circuitous route to Mars, because it could not achieve the velocity needed to travel to the planet directly. "Being able to launch a robust spacecraft, overcoming minor glitches, sending it towards Mars innovatively, braving adverse space weather and radiation hazards while maintaining reliable communication for nearly ten months — it is already a significant achievement for India's first deep-space endeavour," says Lele. ■

Additional reporting by **Alexandra Witze**.

GUY CALAF/POLARIS/EYEVINE

Loss of crops to drought contributed to a food crisis in Ethiopia in 2008.

AGRICULTURE

Cross-bred crops get fit faster

Genetic engineering lags behind conventional breeding in efforts to create drought-resistant maize.

BY NATASHA GILBERT

Old-fashioned breeding techniques seem to be leading genetic modification in a race to develop crops that can withstand drought and poor soils.

As the climate warms and rainfall becomes more erratic, farmers worldwide will increasingly need crops that can thrive in drought conditions. And the high costs of fertilizers — along with the environmental damage they can cause — are also pushing farmers to look for crop varieties that can do more with less.

The need for tougher crops is especially acute in Africa, where drought can reduce maize (corn) yields by up to 25%. The Drought Tolerant Maize for Africa project, which launched in 2006 with US$33 million, has developed 153 new varieties to improve yields in 13 countries. In field trials, these varieties match or exceed the yields from commercial seeds under good rainfall conditions, and yield up to 30% more under drought conditions.

An analysis published earlier this year reported that by the project's end in 2016, the extra yields from drought-tolerant maize could help to reduce the number of people living in poverty in the 13 countries by up to 9% (R. La Rovere *et al. J. Dev. Areas* **48,** 199–225; 2014). In Zimbabwe alone, that effect would reach more than half a million people.

The project's success is due in large part to its access to a large seed bank managed by one of its partners, the International Maize and Wheat Improvement Center (CIMMYT) in Mexico City. Breeders from CIMMYT and the International Institute for Tropical Agriculture in Ibadan, Nigeria, searched the collection for maize varieties that thrive in water-scarce regions. The researchers cross-bred these varieties and then mated the most drought-tolerant of their offspring. Several cycles of this process led to seed that was better adapted to water-scarce conditions. In a final step, project scientists cross-bred these plants with varieties that have been successful in Africa.

"It is a painstaking and expensive process," says Kevin Pixley, director of CIMMYT's genetic resources programme.

The CIMMYT researchers established that certain characteristics predict how a maize plant will fare in drought. One of the most telling is the number of days between when the plant's male organs shed pollen and when the female silks emerge. When water is scarce, the silks emerge late. If the delay is long enough, they emerge after the plants have released their pollen and are not fertilized.

"Finding out this relationship was very important to be able to select for drought tolerance," says Pixley. By favouring plants with shorter intervals between pollen release and silk emergence, breeders were able to produce maize that was more resistant to drought.

Drought tolerance is a complex trait that involves multiple genes. Transgenic techniques, which target one gene at a time, have not been as quick to manipulate it. But CIMMYT and six other research organizations are also developing genetically modified (GM) varieties of drought-resistant maize, in collaboration with agricultural biotechnology giant Monsanto in St Louis, Missouri. Coordinated by the African Agricultural Technology Foundation in Nairobi, the Water Efficient Maize for Africa project aims to have a transgenic variety ready for African farmers by 2016 at the earliest.

Like drought resistance, maize's ability to grow in nitrogen-poor soils is genetically complex, and the need for varieties that do well with little fertilizer is pressing. Most African farmers can afford only one-tenth the amount of fertilizer recommended for their crops. This is one of the biggest problems they face, says Biswanath Das, a maize breeder at CIMMYT.

Researchers at CIMMYT are working to address that problem through the Improved Maize for African Soils (IMAS) project, a collaboration with the Kenya Agricultural Research Institute in Nairobi; the South African Agricultural Research Council in Pretoria; and DuPont Pioneer in Johnston, Iowa. The 10-year, US$19.5-million project is pursuing conventional and transgenic approaches.

Since its launch in 2010, IMAS has developed 21 conventionally bred varieties. Over the next year the project's leaders hope to commercialize these varieties and introduce them in eight countries. In field tests, IMAS varieties yielded up to 1 tonne per hectare more in nitrogen-poor soils than did commercially available varieties. By contrast, the project's researchers say that they are at least 10 years from developing a comparable GM variety.

Conventional breeding will probably have a greater impact, says Das, "but it is important to consider all options". ■

CORRECTIONS

The News Feature 'Survival of the fittest' (*Nature* **513,** 157–159; 2014) referred to the wrong Possession Island. The penguin work was done on the French Base d'Alfred Faure in the Crozet archipelago.

The World View by Casparus J. Crous (*Nature* **513,** 7; 2014) implied that Saudi Arabian scientists on highlycited.com were all at a single university. In fact, most were at one institution but several came from three other universities in Saudi Arabia.

Targeted culling and hunting contests can control lionfish populations.

ALEX MUSTARD/NATUREPL.COM

BOUNTY HUNTERS

Destructive lionfish are invading coral reefs in the Americas, but fishing competitions can help to keep the problem species in check.

BY HANNAH HOAG

Stephanie Green plunged her hands, sheathed in thick black gloves, into a cooler full of lionfish. Skilfully avoiding its 18 venomous spines, she plucked one out and laid it on a table to record its length. Nearby, volunteers were chopping up the brown, red and white striped fish to make ceviche and passing the dish into the crowd.

As they nibbled on the food, teams of scuba divers milled around the scoring area. They checked out each other's catches and argued over who would be taking home the more than US$3,500 worth of prizes from the 2013 lionfish hunting derby in Key Largo, Florida.

"At the check-in time it's a mad rush, with teams coming in with coolers of fish, trying to beat the clock," says Green, the chief scientist of the contest and a marine ecologist at Oregon State University in Corvallis. By the end of that day last September, Green and the other scorekeepers had counted 707 lionfish, from one smaller than a golf ball to one that stretched nearly two soccer balls long.

The hunting competition is part of an effort to tackle an invasive species that has been identified as one of the world's greatest conservation issues[1]. Since lionfish (*Pterois volitans*) first appeared on the eastern seaboard of the United States in the 1980s, the voracious predators have gobbled up coral-reef fish from North Carolina to Venezuela. Officials responsible for protecting reefs have struggled to find ways to control populations, and managers are embracing these fishing contests in a handful of coastal communities.

The strategy is a bit of a gamble, given that competitions to catch other invasive species — such as pythons in Florida — have had limited success. But the data collected by Green show that even one-day contests can effectively knock down local populations. Her findings and those from other hunting efforts offer lessons on how a little bit of reward money — coupled with science and outreach — can help to keep invasive species in check. "We can't control lionfish in the entire ocean, but derbies can have high impacts locally," says James Morris, an ecologist with the US National Oceanic and Atmospheric Administration in Beaufort, North Carolina.

Like many invasions, the lionfish conquest started small. The fish are normally found in the western Pacific Ocean, Indian Ocean and Red Sea, where predators and competitors keep the populations under control. Genetic analysis[2] suggests that roughly a dozen fish were first introduced off the Florida coast, either accidentally or intentionally released from aquariums. From there, the population exploded. Lionfish spawn almost continuously, releasing 2 million eggs a year, and they have few predators or competitors in their new home.

"At first people thought they were funny, beautiful," says Mark Vermeij, a conservation biologist at the Caribbean Marine Biological Institute on the island of Curaçao. But opinions changed as the lionfish took over, he says. "Quite quickly they were everywhere. They were like cockroaches."

Since they were first spotted near Fort Lauderdale, Florida, in 1985, lionfish have colonized more than 4 million square kilometres — throughout the Caribbean Sea, the Gulf of Mexico and all along the Atlantic coastline of the southern United States — and show no signs of relenting. Marine ecologists worry that the invasion will eventually extend to Uruguay, stopped only by winter water temperatures. It could become one of the most ecologically harmful fish introductions in the western Atlantic, says Mark Hixon, a marine ecologist now at the University of Hawaii at Manoa, and Green's supervisor at Oregon State. At some sites off the coast of North Carolina and in the Bahamas, the populations are 5–15 times denser than in the fish's natural range, sometimes even reaching 400 fish per hectare.

"QUITE QUICKLY THEY WERE EVERYWHERE. THEY WERE LIKE COCKROACHES."

The conquest could have profound effects on the biodiversity of coral-reef ecosystems. Lionfish consume whatever fits in their maws — and a lot of it. A DNA analysis[3] of the stomach contents of 157 lionfish caught in the Mexican Caribbean identified 43 crustacean and 34 fish species, including parrotfish, French grunt and graysby — important sources of food for local people. Without natural predators, a lionfish can gobble up 79% of the juvenile fish on a reef in as little as five weeks.

The feeding frenzy could also lead to larger problems. Some of the fish they prey on clean algae off coral reefs and are already overfished in the Caribbean. Without these essential species, algae could outcompete the corals. Simulations by Jesús Ernesto Arias-Gonzàlez at the Center for Research and Advanced Studies of the National Polytechnic Institute, in Mérida, Mexico, have shown that a lionfish invasion would decrease the biomass of corals in a Caribbean reef by about 10% within ten years[4].

OUT OF CONTROL

Green did not set out to study lionfish. She had just started a doctorate in conservation biology when she travelled to the Bahamas in 2008 with her adviser Isabelle Côté, a biologist at Simon Fraser University in Burnaby, Canada. A student they visited was seeing lionfish all over her study sites. "Nobody knew anything about them, the basics of where they were, or what they ate," says Green.

Green and Côté wondered whether the native fish would return if they removed the lionfish. In December 2009, they staked out 24 patches of reef and arranged for scuba divers to prune the population of lionfish at the sites every month for 18 months. They predicted that the culling effort would need to remove 25–92% of the predators, depending on the site, to keep them from consuming too much of the prey species. By the end of the experiment, the native fish had rebounded by 50–70% in the reefs that reached the targeted level of protection[5].

Green and Côté were not the only ones to hunt down lionfish. Earlier that year, the Reef Environmental Education Foundation (REEF) in Key Largo, Florida, had started running derbies in the Bahamas to increase local awareness of the invasion. Green, who had been collaborating with the foundation during her PhD, got involved in planning the first hunts.

She later decided to use the competitions to test whether limited hunts could have an impact. With the help of volunteers outfitted in scuba and snorkels, Green counted lionfish at 60 sites before and after the derbies in Key Largo and the Bahamas in 2012 and 2013. On the basis of a preliminary analysis of the derbies, she says, "there were dramatic drops in the densities of lionfish in the sites where people fish." After the competitions, lionfish densities were slashed by more than 60% over a 100–150 km^2 area compared with pre-derby levels. "It's like pulling weeds from your garden," she says. "You're not going to completely get rid of them, but below a certain level, they won't cause problems."

Lionfish recolonized the sites within six months, but the animals were significantly smaller, which helped to reduce pressure on the reef. Smaller lionfish eat less, prey on smaller fish and produce fewer young.

Ted Grosholz, a marine ecologist at the University of California, Davis, says that the data collected by Green and REEF support the idea that derbies can effectively control lionfish populations in selected areas. They also dovetail with results from other lionfish control efforts. When the fish invaded the Dutch Caribbean in 2009, volunteers immediately began to use spear guns to remove lionfish from the island of Bonaire, but did nothing in neighbouring Curaçao. After two years of spearfishing, Vermeij and his colleagues found that the lionfish biomass in the treated areas of Bonaire was just one-third of that in the unfished areas, and about one-quarter of what was seen in Curaçao[6].

ON TARGET

The lionfish contests have been much more successful than some other efforts that have used hunters to control invasive species. In 2013, for example, the Florida Fish and Wildlife Conservation Commission organized the first Python Challenge, a month-long event with cash prizes that enlisted professional and amateur hunters to remove Burmese pythons (*Python bivittatus*). But the pythons proved tough to catch because they are hard to spot in the Florida brush; the hunters caught just 68 snakes from a population that is estimated at 30,000–100,000.

Jason Goldberg, a biologist at the US Fish and Wildlife Service in Arlington, Virginia, says that derbies could be improved by incorporating the results of research. Organizers need to calculate how many individuals to remove, whether it is better to cull older or larger individuals and how their density affects the health of the population. That information can then be used to set hunting targets — and prevent the kinds of problems that arose when Australia culled red foxes (*Vulpes vulpes*). The 2002–03 Victorian Fox Bounty Trial removed one-fifth of the state's red foxes but ended up boosting the population because the survivors thrived when they had less competition for food[7].

Cash incentives can help by drawing amateurs into efforts to control invasives. In the Pacific northwest, for example, anglers are offered $4–8 for every northern pike minnow (*Ptychocheilus oregonensis*) they capture to deter the fish from preying on young salmon. The programme has removed more than 3.9 million fish and slashed predation by 40%.

Goldberg says that research on lionfish derbies should offer insight into how often — and when — they should take place at each location. He adds that new steps might be needed, such as encouraging commercial fishing of lionfish to make the species more common in restaurants.

The lionfish invasion and the success of the derbies has led to policy changes in Florida. In August, wildlife regulators relaxed hunting restrictions in the state to allow divers wearing rebreathers — devices that allow them to remain in the water for longer — to harvest lionfish. It will also now allow derby participants to spear lionfish in areas where spearfishing is otherwise prohibited. "For marine protected areas to function as conservation areas, it's important that the biology and ecology be conserved to the highest level possible, and that now requires lionfish control," says Morris.

With her results pointing in a positive direction, Green intends to continue analysing data from lionfish derbies, including an event in Key Largo on 13 September. When she shares her research findings with the divers, it tends to fire them up, she says. "There's this good community feeling at the derbies that this is a tool that can have a positive effect and help to suppress the invasion." ■

Hannah Hoag *is a freelance writer in Toronto, Canada.*

1. Sutherland, W. J. *et al. Trends Ecol. Evol.* **25,** 1–7 (2010).
2. Freshwater, D. W. *et al. Mar. Biol.* **156,** 1213–1221 (2009).
3. Valdez-Moreno, M., Quintal-Lizama, C., Gómez-Lozano, R. & García-Rivas, M. del C. *PLoS ONE* **7,** e36636 (2012).
4. Arias-González, J. E., González-Gándara, C., Cabrera, J. L. & Christensen, V. *Environ. Res.* **111,** 917–925 (2011).
5. Green, S. J. *et al. Ecol. Appl.* **24,** 1311–1322 (2014).
6. de León, R. *et al. Endang. Species Res.* **22,** 175–182 (2013).
7. Pasko, S. & Goldberg, J. *Manag. Biol. Invasions* (in the press); corrected proof available at http://go.nature.com/o7drth.

ILLUSTRATION BY DENIS CARRIER

PRIDE IN SCIENCE

The sciences can be a sanctuary for gay, lesbian, bisexual and transgender individuals, but biases may still discourage many from coming out.

BY M. MITCHELL WALDROP

"I was the golden child," says Justin Trotter, thinking back to his teenage years living near the Kennedy Space Center in Brevard County, Florida. The handsome, articulate son of a devout Mormon family, he earned top grades, assembled winning projects for science fairs and worked in university laboratories from the age of 14.

But he was also wrestling with a secret. Trotter, now a neuroscience postdoc at Stanford University in California, says that as early as the ages of 11 or 12 he had begun to sense that he was attracted to boys — a feeling that he had always been taught was shameful. So all through his teens and early twenties, he says, he struggled to keep his sexuality hidden, to appear masculine, to blend in.

"I dreaded dealing with it," says Trotter. By his undergraduate years at university he was suffering from exhaustion, depression and panic attacks. "My only escape was to work in the lab," he says: "It was my haven." But the stress took its toll even there. "I felt my memory wasn't good. I wasn't as sharp as I could be."

It was not until the last two years of his graduate studies, at the University of South Florida in Tampa, that Trotter finally came out, confiding to a few close friends that he was gay. As the word spread, he found his depression lifting. His energy improved. His work became more focused.

"When I felt I could just be who I am, a full person," says Trotter, "then it was definitely good for the science."

That message is being heard in more and more laboratories and research centres around the world. People who identify as lesbian, gay, bisexual or transgender (LGBT) have long faced discrimination or worse: they are still considered outcasts or even outlaws in most Muslim nations, as well as in Russia and parts of Asia. But attitudes are changing. According to a survey published last year by the Pew Research Global Attitudes Project, openly gay individuals have high levels of public acceptance across broad swathes of Western Europe, Australia, Canada and Latin America (see 'Degrees of acceptance'). Nowhere is this change more visible than in the United States, home of the world's largest research enterprise, where public attitudes are shifting towards acceptance of LGBT people faster than in almost any other nation. Courts and legislatures are lifting restrictions on same-sex marriage in state after state, often in the face of vehement opposition from social conservatives, and LGBT equality has emerged as a dominant civil-rights issue.

"This is an important time in history for the LGBT community," says Trotter — not unlike the period several decades ago when women and under-represented ethnic minorities began their push for greater recognition in science. Just as those groups once did, LGBT researchers are trying to seize the moment by creating an infrastructure of organizations and interest groups geared towards helping one another with information, support and networking (see *Nature* **505,** 249–251; 2014).

OUT IN THE OPEN

In this newly open environment, LGBT scientists are finding it easier to declare themselves — or at least, to think about doing so. "I'm getting a constant stream of e-mails from young scientists: 'Can I meet with you?'," says Ben Barres, a Stanford neuroscientist who transitioned from female to male in 1997, and who has become a prominent spokesman for LGBT issues in science.

But just as for ethnic minorities and women, there is still a long way to go. Many LGBT scientists fear coming out — if only because publications, career progression and promotion are based heavily on the judgement of fellow scientists, which might be influenced by conscious or unconscious bias. And many students may be avoiding a research career entirely — although no one knows, because no one has counted.

"I worry that there is a vast pool of talent that might be being lost to science," says Trotter. The only way to change that, he says, is for the scientific community to reach out to its LGBT members, and have an honest conversation.

The lab can be an excellent place for that dialogue, says Kale Edmiston, a neuroscience graduate student at Vanderbilt University in Nashville, Tennessee. "The cool thing about scientists is that we try to withhold judgement and gather information," he says. That is exactly what happened when Edmiston began transitioning in 2010. He told everyone in his research group that he would be taking hormones and that his appearance would change. They responded with a combination of sympathy, interest and curiosity. "A lot of my peers and colleagues have really listened and heard what I'm saying," says Edmiston. Rachael Padman had much the same experience in the early 1980s, when she transitioned from male to female while a graduate student in astrophysics at the University of Cambridge, UK. "One colleague was never able to say 'she' instead of 'he'," says Padman. "But that was just one in the last 35 years. From almost everyone else, I just never get any vibes at all."

The mores of research can work the other way, too, says Vivian Underhill, a field hydrologist with the University of Colorado Boulder's Institute for Arctic and Alpine Research, and author of 'Queered Science': a blog series about LGBT researchers. "As scientists we like to think that we're objective," she says — that personal and social issues should be kept separate from the real work. "And mostly that's a good thing," she says, but too often it leads people to assume that they can eliminate biases by not talking about them. "That just enables the fear to propagate."

> "WHEN I FELT I COULD JUST BE WHO I AM, A FULL PERSON, THEN IT WAS DEFINITELY GOOD FOR THE SCIENCE."

It can also make it hard to comprehend the stark loneliness that comes with being LGBT in a majority-straight world. Unlike women or ethnic minorities, LGBT people are not automatically born into a peer group, says Darrin Wilstead, director of operations for the Point Foundation, an LGBT scholarship and mentoring fund based in Los Angeles, California. Almost always, he says, "they come home to a family that does not share their identity" — and may not understand or accept it. Everyone has to come to terms with their sexuality as they grow up. But LGBT individuals often have to begin that journey in isolation.

Gay, lesbian and bisexual feelings typically emerge around the time of puberty — although they may start much earlier. "When you're a kid, you just aren't as aware of who you're attracted to," says Eli Capello, an undergraduate neuroscience major at Centenary College of Louisiana in Shreveport. Transgender issues, by contrast, can become obvious at a very early age. "I knew something was up when I was 3 and 4," says Capello, who transitioned when he was 18. "I just didn't know what."

GROWING PAINS

Many people lack basic knowledge about gender identity, which is different from sexual orientation. The latter concerns who a person is attracted to; gender identity is about the body someone is born into and whether it matches what the brain is insisting. Either way, LGBT individuals typically find themselves struggling to deal with all this in their teens and early twenties — precisely when science students are also supposed to be mastering their fields. Some respond by throwing themselves into their coursework. "College was a great place to distract myself," says Underhill, who did not tell close friends that she was lesbian until just before her graduation. "I didn't even allow myself to look for online sources that might have been helpful, because I didn't want to go there."

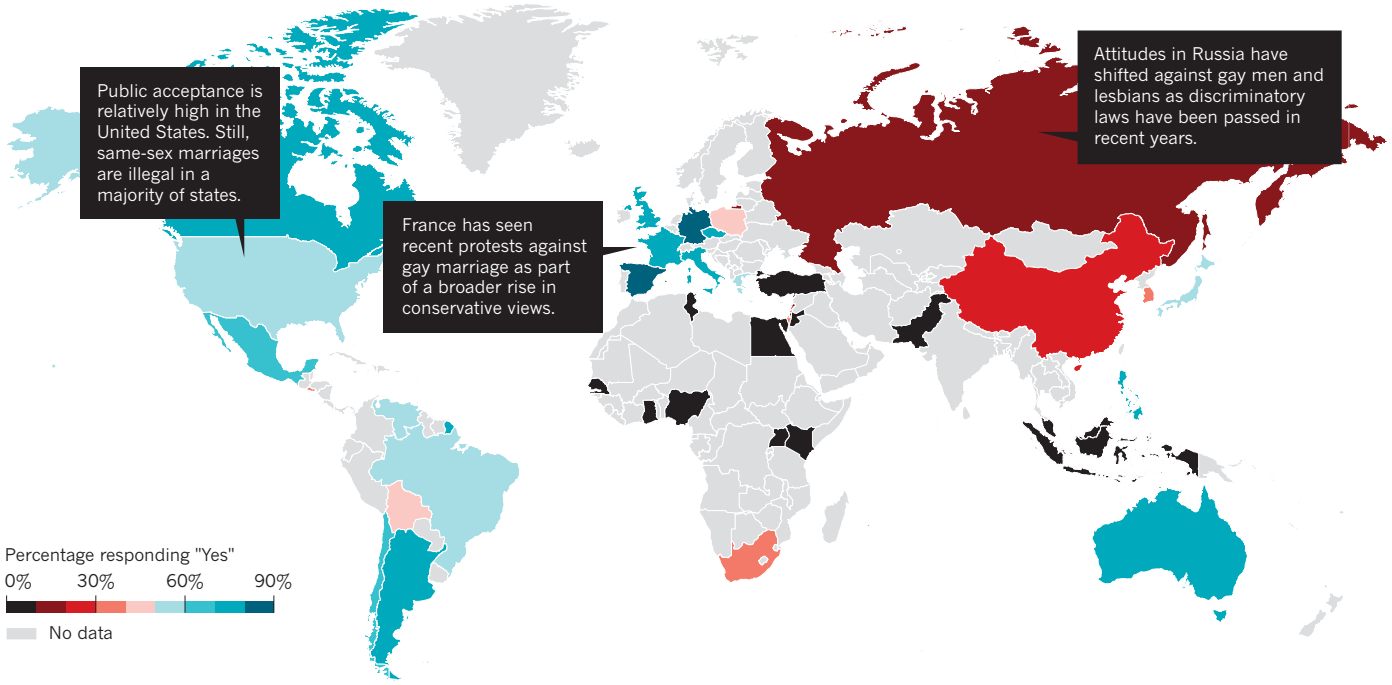
But the kind of emotional turmoil that Trotter describes is also very common. According to the US Centers for Disease Control and Prevention in Atlanta, Georgia, gay, lesbian and bisexual teenagers generally experience high levels of bullying and drug abuse, and are more than twice as likely to attempt suicide as their heterosexual peers. The lack of data means that there is no way of knowing how often this leaves promising students too stressed to attempt challenging science, technology, engineering or mathematics (STEM) degrees. But anecdotal evidence suggests that it does happen. At the Point Foundation, says Wilstead, "we found out that some areas of studies, like law, medicine and STEM, were much more challenging" for scholarship recipients trying to maintain their grades. And that, in turn, may help to explain why only about 10% of the foundation's applicants were in STEM fields until five years ago, when the foundation began vigorous outreach efforts that helped to raise the fraction to around 20%.

Adding to the loneliness and stress is the highly charged decision of whether to come out at all. To stay in the closet perpetuates the

DEGREES OF ACCEPTANCE

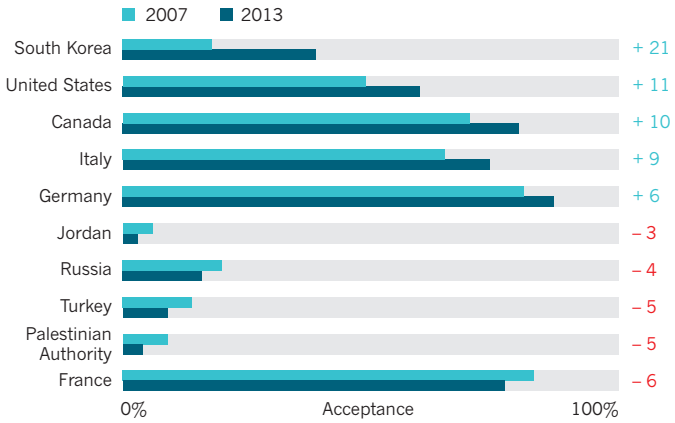
In 2013, as part of a survey on global attitudes, the Pew Research Center in Washington DC asked people in 39 countries: "Should society accept homosexuality?" (It did not ask about transgender people.)

SOURCE: *THE GLOBAL DIVIDE ON HOMOSEXUALITY* (PEW RESEARCH CENTER, 2013); AVAILABLE AT GO.NATURE.COM/DJRZTA



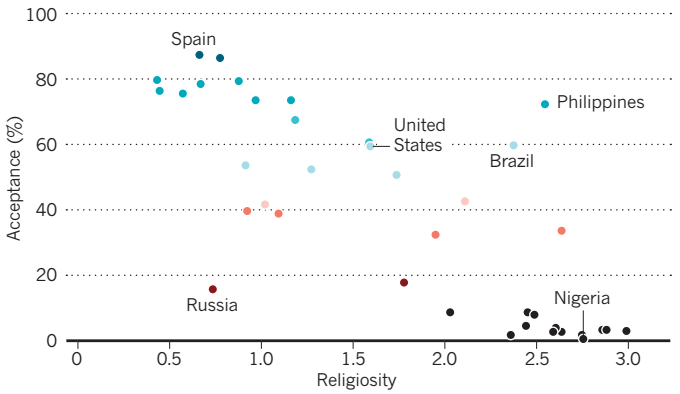
STABLE VIEWS – MOSTLY

Opinions had not changed much since 2007. But there were some notable exceptions: South Korea, the United States and Canada each increased acceptance by ten percentage points or more.



TOLERANCE AND RELIGION

The Pew data show that tolerance tends to be low in countries with high 'religiosity', a measure of how much importance people place on prayer, ritual and belief. Countries with low religiosity tend to have high acceptance.



turmoil, but the consequences of not staying there can be awful. After Capello came out as transgender at age 16, for example, his relationship with his family deteriorated to the point at which he had to leave home and attend a boarding school paid for by his grandmother.

Jun Ding, a Stanford neuroscientist who grew up in Shanghai, China, before moving to the United States for his PhD relates a different kind of experience. Although Chinese laws that were used to discriminate against gay people were repealed in 1997, there are no protections and very little public discourse on the subject. Ding says that his parents still do not understand when he tries to explain that he is now living with his husband, whom he married under California law. "It's not like the US, where almost every movie has a gay role," he says. "In China, a lot of people just don't have the concept of gay life."

PEER PRESSURE

Even if family rejection is not a concern — and it seems to be less common than it once was, says Wilstead — the decision is not necessarily any easier. More than in almost any other field, a researcher's career is based on peer review in the widest sense, says Eric Patridge, a chemist at Yale University in New Haven, Connecticut, and president of Out in Science, Technology, Engineering, and Mathematics, a national LGBT student group. Colleagues' opinions weigh heavily when it comes to funding, collaboration, publication, hiring, promotion and almost every other decision. In a highly competitive environment, every LGBT researcher has to worry that coming out will trigger unconscious biases that could ruin his or her chances. Studies from the US National Institutes of Health over the past few years suggest that such bias may well be a problem for other minorities (see *Nature* **512,** 243; 2014), and there is no reason to think that LGBT researchers are exempt.

That may be why Barres hears from many young LGBT scientists who are afraid to come out, even in the San Francisco Bay Area of California, historically one of the most tolerant regions in the United States. And it is worse in the more socially conservative

regions of the country: "When I talk to people from down south, the fears are so much stronger," he says.

Those fears are justified, says Trotter. After he came out during his graduate studies in Florida, he says, some of the more religious, socially conservative students in his research group became noticeably reticent around him. "The scientific community is still comprised of people at varying stages of social progress in ideas," he says — a reality that he has not escaped. He is very comfortable at Stanford, he says. But once his postdoc appointment there is over, he may well have to apply for tenure-track jobs back in Florida or another less-tolerant part of the country — "where my ability to acquire tenure or run a successful research programme would be in question".

And that is just in the United States. Scientists and science students seem to be equally reticent in LGBT-friendly Western Europe — and even more so in China. In many parts of the Middle East and Africa, moreover, LGBT activity is punishable by law — in some cases, with execution (see *Nature* **509,** 274–275; 2014). "So if you're a chemist or geologist for an oil company, you'd better be in the closet if they send you to one of those countries," says Rochelle Diamond, who chairs the board of directors of the National Organization of Gay and Lesbian Scientists and Technical Professionals in Pasadena, California.

DIFFICULT CONVERSATION

Coming out can be even more daunting for transgender people — especially if the announcement coincides with the start of sex-reassignment treatments. Added to the emotional stress and professional concerns are the physical effects of interventions such as taking hormones. "There's a reason you're supposed to go through puberty before you get to college," says Capello. Kate Forbes, who now works for a medical information-technology company, transitioned while she was earning her doctorate in ecology at the University of Wisconsin–Madison; she describes nights curled up in pain on her futon after electrolysis treatments to get rid of the hair on her legs.

Then there are the awkward conversations. "Every time I'd start a course I would have to have a very personal discussion with the professor about things like male pronouns," says Lucas Cheadle, a neuroscience postdoc at Harvard University in Cambridge, Massachusetts. Making the situation doubly difficult was that he transitioned while he was an undergraduate at Smith College — a women's university in Northampton, Massachusetts. "I missed out on a lot of mentorship relations because of the difficulty of explaining," he says.

Even ordinary paperwork can become a major burden. Newly transitioned individuals can spend endless hours struggling to convince sceptical bureaucrats that they have a legitimate claim to their own university transcripts, publication lists, birth certificates, driving licences, credit cards and more — all now in a different name. Even in tolerant Western Europe, officials may demand extensive documentation before making the required changes, says Padman. "Britain is quite unusual in that respect," she adds: the Gender Recognition Act 2004 declares that each individual is the gender he or she decides to be. "They can get a new birth certificate with the new gender, and have all the legal rights of the acquired gender." Backing that up is the Equality Act 2010, which forbids discrimination again trans people. "It discourages the tabloids from making a huge deal about someone's sex reassignment on the front page, the way they used to," says Padman.

> "THE SCIENTIFIC COMMUNITY IS STILL COMPRISED OF PEOPLE AT VARYING STAGES OF SOCIAL PROGRESS IN IDEAS."

The United States has certainly not gone as far towards recognition as Britain. But even so, the situation seems to be improving rapidly for younger LGBT people — not least because of the Internet. "When Point was founded 13 years ago," says Wilstead, "our scholars were saying that they couldn't find anyone who was older, gay and successful." But today, thanks to Facebook, Twitter and the plethora of other social-media sites, it is much easier to make such connections. And one consequence, says Wilstead, is that LGBT people are coming out much earlier than they used to. Jack Andraka, for example, was a 15-year-old student in Crownsville, Maryland, when he won the grand prize at the 2012 Intel International Science and Engineering Fair, for discovering a test for pancreatic, ovarian and lung cancer. He had come out as gay when he was 13.

Another consequence is an increased sense of solidarity in the LGBT community itself. The group was once defined only by what its members are not: straight. "Aside from that," says Underhill, "my experience as a white lesbian woman may have little in common with that of a black gay man." That is a point echoed by many others: their 'community' is often riven by the same fault lines as the society around it, with gay white men dominant, and women, bisexuals and ethnic minorities each feeling marginalized in their own ways. And only in the past five years or so have transgender individuals begun to become more visible.

But the rising generation tends to be much more concerned about inclusion, says Wilstead. Witness the embracing of the once-derogatory term 'queer'. "It's an umbrella term," he says. "It's basically saying, 'I am just different'." That openness and solidarity, in turn, is making it much easier for young scientists to find mentors and role models. It is impossible to overemphasize how important that is, says Trotter — "seeing that it really does get better, seeing gay and lesbian scientists who have been through it and made it".

But the scientific establishment could give a lot more help in promoting role models than it has so far, say LGBT activists. Barres thinks that the US National Academy of Sciences missed a golden opportunity when he was elected to membership last year. "I asked if they would include in the announcement that I was the first transgender scientist to be elected," he says. "They never did." (Electrical engineer Lynn Conway was elected to the parallel National Academy of Engineering in 1989, although she did not come out as transgender for another decade.)

The scientific establishment could also do a lot more about collecting basic data. For example, the US National Science Foundation, which compiles detailed statistics about women, under-represented minorities and the prevalence of various disabilities among US researchers and STEM students, does not currently ask about LGBT identification. Nor do there seem to have been systematic, large-scale studies of the social environment for LGBT researchers. How much stress do they really feel if they stay closeted in the lab? What are the actual effects on their health and productivity? And if they do come out, are they really less likely to be funded, hired or promoted? At least one team — sociologists Erin Cech of Rice University in Houston, Texas, and Tom Waidzunas of Temple University in Philadelphia, Pennsylvania — is hoping to carry out a survey of 2,000–3,000 LGBT scientists and engineers, but has yet to get funding.

Without such data, says Trotter, it is impossible for the funding agencies to know whether LGBT people are over- or under-represented in the research fields, whether there is a need for more support programmes and counselling, or whether they should offer special fellowships for young LGBT researchers in the way they now do for women and minorities. "We don't have numbers," says Trotter, "and that's frustrating for us as scientists."

Still, without minimizing the challenges that remain, older LGBT scientists stress how far the world has come in a remarkably short time. "When I'm contacted by young people," says Barres, "I always tell them that the fears are so much greater than the reality. And I always encourage them to be open, because they will be so much happier. If you're doing good science, if you're a great teacher — that's what matters." ■

M. Mitchell Waldrop *is a features editor for* Nature *in Washington DC.*

COMMENT

ED KASHI/VII/CORBIS

Common asthma drugs can work less well for children of some ethnicities.

Missing patients

Effective clinical studies must consider all ethnicities — exclusion can endanger populations, says **Esteban G. Burchard**.

In 1997, when I was a resident at Harvard Medical School in Boston, Massachusetts, an African American teenager was found dead just blocks away from the teaching hospitals. He had died of an asthma attack while clutching his inhaler.

It is widely known that racial and ethnic minorities in the United States have higher rates of diseases such as asthma[1] and cancer[2], and receive worse care[3]. Compared with white people with similar conditions, minority individuals get fewer heart bypasses and influenza vaccinations.

Less well known is the fact that many drugs work better in people of European origin than in others. One class of asthma drug (long-acting β2-agonists) is even associated with higher mortality in African Americans[4].

Populations of non-European descent are harmed because they are not studied as intensely, and clues that could reveal new aspects of disease biology are missed. Including diverse populations in clinical and biomedical research is a must, ethically and scientifically. Research infrastructure needs to be retooled accordingly.

My mother was Mexican, an overworked single parent who learned English and put herself through university. I spent much of my time growing up with a Chinese surrogate family. My wrestling coach, an African American and member of the 1984 US Olympic team, became my mentor and father figure. Later, in medical school, I lived in housing that was set up by Jewish students. These experiences have prompted me to consider health disparities across racial and ethnic populations, which I discuss here using the terms and criteria established by the US Centers for Disease Control and Prevention (CDC; see go.nature.com/a2euvo).

The year that the young man in Boston died, my colleagues and I identified a variant associated with asthma in the gene for interleukin 4, a cell-signalling protein that coordinates immune and inflammatory responses. In our study[5] of 772 individuals, the variant was associated with lower levels of lung function, leading to more ▶

severe disease in white people. Although black children are more likely than white children to be diagnosed with and die from asthma (see 'Asthma inequalities'), few black patients were included in the study, so we had insufficient statistical power to establish the genetic association in black people. However, our analysis found that the variant was 40% more common in black people. This led me to wonder whether some health disparities might result from genetic differences, as well as social and environmental factors.

That same year, the CDC published data showing that the occurrence of and deaths from asthma were threefold higher in Hispanic communities in the northeastern United States than in those on the west coast. I immediately thought that the observation could result from genetic differences between Puerto Ricans (concentrated in the east of the country) and Mexicans (concentrated in the west). This realization spurred the Genetics of Asthma in Latino Americans (GALA) study, which started in 1998 in Boston, New York and San Francisco, California. For one analysis, children with asthma were asked to breathe into a measuring apparatus after receiving standard treatments. The research[6] showed that the biggest predictor of drug response was ethnicity — stronger than age, sex or disease severity. Commonly prescribed asthma medications worked less well for Puerto Ricans than for Mexicans and African Americans.

"Grant applications should be regarded more favourably, not less, for analysing minority populations."

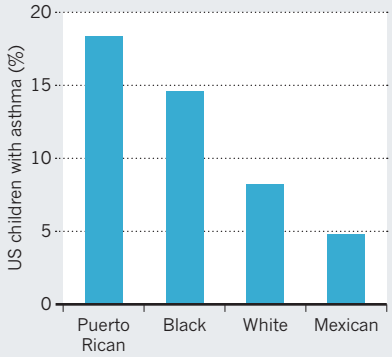
Such disparities occur across other ethnicities and conditions. Heart disease and stroke are the two leading causes of death worldwide, and the blood thinner clopidogrel is widely prescribed to people who have had a heart attack or a stroke. In March, officials in Hawaii sued the drug's manufacturer for failing to disclose that it is frequently ineffective in the state's largely east Asian and Pacific Islander population, placing them at higher risk of recurrent heart attacks.

VICIOUS CYCLE

Many barriers keep studies that could characterize such disparities from being proposed, funded, carried out and published. The hospital where I work runs dozens of clinical trials, but it serves mainly people of European and Asian descent. To recruit enough subjects for the GALA study we sent staff to other parts of the San Francisco Bay Area, to Mexico and to Puerto Rico. We established a network of physicians with experience serving diverse communities, used bilingual clinical coordinators and engaged community clinics, religious leaders and community activists. That I am a bilingual physician–scientist from an ethnic minority was invaluable for brokering these connections.

ASTHMA INEQUALITIES

Genetic variants contribute to disparities in asthma incidence and treatment efficacy.



SOURCE: REF.1

Once participants were recruited, we measured genetic ancestry using reference data from the 1000 Genomes Project and the Human Genome Diversity Project. This was not easy — fewer than 4% of genetic association studies have been conducted in people of non-European descent. We had to create our own human genetic reference panels by teaming up with another investigator who had collected samples from Native Americans.

Our work paid off. We were able to show that considering genetic ancestry can improve the accuracy of diagnosis of lung disease in African American and Mexican populations. We have also identified genetic variants that might explain why asthma drugs work less well for Puerto Rican and African American children. Clinical trials are now under way to assess the efficacy of asthma medications in different ethnic populations, based on genetic variants.

Such work, focused on minority populations, faces a vicious cycle. As a reviewer for the US National Institutes of Health (NIH), which is funded by US taxpayers, I witnessed how grant applications that propose genetic analyses in minority populations in the United States are criticized because reviewers considered these populations more difficult to analyse than more-genetically-homogenous European populations. Sadly, I believe that many NIH reviewers see rich genetic ancestry largely as a potential confounder. They do not appreciate that it can be leveraged to reveal new risk factors.

Publishing such results is also difficult. Most high-impact journals require that an association be found in samples from two independently recruited studies. This demand is straightforward in European populations, because many banked samples exist. It is much harder to meet for other groups.

Disparities are self-perpetuating. Minority scientists are often best placed to gain community 'buy in' and trust in minority populations, but these scientists are at a disadvantage in other ways. According to one analysis, black scientists in the United States were 13% less likely to get NIH funding than white researchers[7]. In short, investigators who want to focus on minorities face extra challenges.

COUNT EVERYONE IN

The NIH Revitalization Act of 1993 mandated that NIH-funded research must include minorities. Twenty-one years later, diversity-focused clinical research is still the exception, not the rule. Although black people and other minorities in the United States have greater rates of and mortality from cancer than white people[2], they are generally less likely to be enrolled in clinical trials. Of the 10,000 clinical trials funded by the National Cancer Institute since 1993, only around 150 studies focused on racial or ethnic minorities[8].

Such gaps and their effect on health care must be assessed. Funding agencies should do more to collect evidence on what research is needed, promote research training, and provide venues for discussion of disparities in biomedical research. At a minimum, the race and ethnicity of study participants should reflect the population with the disease being investigated. Grant applications should be regarded more favourably, not less, for analysing minority populations. Journals should require appropriate representation and analyses before publishing clinical studies.

Investigators must also form partnerships with physicians and residents in under-represented communities — they too have a vested interest in improving studies. Finally, there must be increased recruitment of minority physicians and scientists and mechanisms to enhance their training and retention.

At every stage of the scientific discovery and review process, investigators should keep in mind that ancestry can contribute to differences in disease and drug response. To do otherwise is to ensure worse health for us all. ■

Esteban G. Burchard *is professor of bioengineering and therapeutic sciences, and of medicine, at the University of California, San Francisco, USA.*
e-mail: esteban.burchard@ucsf.edu

1. Moorman, J. E. *et al. National Surveillance of Asthma: United States, 2001–2010* (National Center for Health Statistics, 2012).
2. Aizer A. A. *et al. Cancer* **120,** 1532–1539 (2014).
3. Durazzo, T. S., Frencher, S. & Gusberg, R. *JAMA Surg.* **148,** 617–623 (2013).
4. Currie, G. P., Lee, D. K. & Lipworth, B. J. *Drug Saf.* **29,** 647–656 (2006).
5. Burchard, E. G. *et al. Am. J. Respir. Crit. Care Med.* **160,** 919–922 (1999).
6. Naqvi, M. *et al. J. Asthma* **44,** 639–648 (2007).
7. Ginther, D. K. *et al. Science* **333,** 1015–1019 (2011).
8. Chen, M. S. Jr, Lara, P. N., Dang, J. H., Paterniti, D. A. & Kelly, K. *Cancer* **120,** 1091–1096 (2014).

ULUTUNCOK/LAIF/CAMERAPRESS

A researcher works in a malaria-drug laboratory in the Democratic Republic of Congo.

Discuss inequality

Confront economic differences to strengthen global research, urge **P. Wenzel Geissler** and **Ferdinand Okwaro**.

Health research in Africa operates across vast economic and political inequalities. These shape collaborating scientists' work and lives, running counter to ideals of equal partnerships. Most scientific staff from 'the south' (as current euphemism circumscribes impoverished countries in sub-Saharan Africa) are employed by local institutions or on fixed-term contracts. Although these researchers have advantages over non-collaborating local colleagues, they have lower salaries than their northern counterparts and smaller allowances for health, housing, retirement or their children's education. Yet when expatriate and local scientific workers depart after the workday for opposite sides of town, which might retain overtones of class and colonial differences, such inequalities go without mention.

This is a problem. Unacknowledged inequalities cause misunderstandings and irritation. Consider young African researchers at an international conference, sleeping and eating outside the five-star venue and missing informal scientific exchanges to save their allowances. To naive colleagues they might seem uncommitted. Or consider a southern principal investigator hired to run part of a multisite trial, who must cover the substantial security, housing and school bills of expatriate colleagues within a limited local budget. As ethnographies of transnational research show, this is not simply the trope of exploited southern scientists and dominating northern ones; it is the reality of global health science[1,2].

We use the term 'unknowing' for the tendency to avoid straightforward talk about obvious inequalities[3]. Unknowing can be about participants' poverty or malfunctioning public health facilities, or be between scientific staff. Discourse is rarely silenced overtly — although speech about inequality is often limited to private chats between peers. Euphemisms — including 'partnership', 'south' and 'north' — deflect attention from material realities. So do practical arrangements: discussing or comparing salaries openly would reveal the systemic challenge of addressing radically different pay rates, so researchers instead negotiate individually for 'reimbursement' or 'per diems' for field days.

The subject is raw. Calling out inequality reveals our limited ability to do the right thing. Northern colleagues in particular sometimes feel that such talk is unfair. As progressive researchers working to improve others' lives, they resent comparisons, even unintended, with exploitative colonialists[4].

In large collaborative sites, open discussion threatens established routines and substantial investments. When disparities seem intractable, talking about them can feel paternalistic, rude, whiny, self-righteous — or pointless. Yet unknowing also subdues essential scientific discussion. The 'normal emergency' state of African health systems often requires research into context-specific tools — adapted reference values for nutrition or toxicity, surgical-safety procedures in understaffed theatres and diagnostic procedures substituting for standard tests[5]. Polite reluctance to acknowledge actual health-care conditions, insisting instead on universal standards and research themes, can produce 'world-class' research that fails to address real-world concerns and opportunities. In turn, this can instil a sense of futility in local scientists[6].

Divergent interests and needs, if ignored, find destructive outlets — such as workplace talk of racism or colonialism, or conflicts fought under the guise of ethical review procedures or even outright legal battles. In short, unexpressed 'us and them' thinking erodes research quality. Frank discussion about money, opportunity and leadership is not, as it sometimes is denigrated, just 'politics'. It is essential.

IT'S GOOD TO TALK

That said, articulating inequalities between unequal parties is difficult. Such conversations may be launched by those who seek redress, but it should also be initiated by leaders or with the aid of social scientists.

This year, we presented fictional cases about everyday research ethics and inequality in a number of sites in east Africa and in European collaborating institutions (see africanbiosciences.wordpress.com). Most participants — especially local collaborators — appreciated the chance to acknowledge inequalities publicly. One presentation initiated the first open discussion about external research in a hospital that had collaborated with a European partner for decades; it yielded suggestions for improved communication and requests for specific diagnostic resources. Even after inconclusive arguments or intense conflicts about the 'realism' of managers and 'radical' demands by junior staff, participants remarked how good it was to talk.

Collaboration under conditions of inequality is not comfortable. Talking about these realities can foster shared goals, more-equal relationships and better science. ■

P. Wenzel Geissler *is professor of social anthropology at the University of Oslo, Norway, and director of research in the Division of Social Anthropology, University of Cambridge, UK.* **Ferdinand Okwaro** *is a research fellow at the Centre of African Studies, University of Cambridge, UK.*
e-mails: p.w.geissler@sai.uio.no; fokwaro@gmail.com

1. Crane, J. *Lancet* **377,** 1388–1390 (2011).
2. Prince, R. J. in *Making and Unmaking Public Health in Africa* (eds Prince, R. J. & Marsland, R.) (Ohio Univ. Press, 2013).
3. Geissler, P. W. *Am. Ethnol.* **40,** 13–34 (2013).
4. Redfield, P. *Cult. Anthropol.* **27,** 358–382 (2012).
5. Feierman, S. in *Evidence, Ethos and Experiment* (eds Geissler, P. W. & Molyneux, C.) (Berghahn, 2011).
6. Lachenal, G. in *Para-States and Medical Science* (Duke Univ. Press, 2014).

Tailor informed-consent processes

The first step in studying mental-health interventions across cultures is to adjust procedures to participants' needs, says **Mónica Ruiz-Casares**.

The enduring mental-health consequences of armed conflict, natural disasters and forced migration are increasingly recognized. But clinicians and the people they help often come from different backgrounds, each unfamiliar to the other[1]. This can result in tensions, inappropriate health services and misleading research.

It is hard to assess how well existing scientific evidence for effective mental-health care applies to any group under-represented in mental-health research. That evidence is based mainly on adult, Western populations. Refugees, migrants and young people from other cultures are much less studied; conducting the necessary research can raise thorny ethical issues. Investigators and their institutional review boards must be flexible and patient so that this work can progress.

Informed, reasoned and voluntary consent is core to the ethical conduct of research, but the norms vary across cultures[2]. Even the standard practice of explaining a research protocol to study participants and then obtaining signed consent can raise complex issues. Some communities value verbal agreements more than written contracts, which they might view with suspicion. The very act of requesting signatures could create mistrust and the misperception that participants are entering into binding agreements that they will not be able to withdraw from. In places with repressive political regimes or for undocumented immigrants, signed consent could put participants in danger. Women who make decisions about their own or their children's participation in research without their partner's consent could face serious family conflict or even violence.

Alternative procedures tailor information and consent to participants' needs. The basis of robust informed consent is respect — for dignity, intercultural dialogue, equality and solidarity. Researchers must find more ways to present information clearly, particularly to non-literate participants. Some investigators use colourful visual aids or multimedia. We are developing visual informed-consent forms with children in Canada[3] and Cameroon in which potential research subjects take photographs to represent ethical issues. Children have, for example, photographed a lock or someone's mouth with a finger crossing it to represent confidentiality.

PETER BARKER/PANOS

Community gatherings can be part of an informed-consent process.

Researchers must also think about potential consequences that a research project might have on participants and their community. This requires time to build trusting relationships and to engage communities and local co-investigators. Focused discussion groups, local advisory committees and participants themselves can share oversight of the appropriateness of study design and implementation. Schedules must allow for presentations to multiple stakeholders and allocate time for participants to reflect on the study and discuss it with others before deciding to participate.

Local beliefs about who has decision-making authority must be considered. In 'collectivistic' cultures, which emphasize the needs of groups over individuals, informed consent operates through relationships. Other community members act as consultants or permission granters[4]. Physicians, elders, senior males or mothers-in-law might be assumed to know best. In western Kenya, community assemblies known as *mabaraza* were consulted in a long-term study of children separated from their parents or orphaned, and they recommended community decision-making in the consent process for biomedical and behavioural research[5].

Informed consent at the community level can conflict with Western standards for voluntary individual decisions. In a child-protection study that I led in Laos, we first obtained explicit permission from provincial, district and village authorities, deemed necessary and sufficient by local standards. But our own principles and those of our institution also made it essential to gain individual informed consent from parents and assent from children. We learned to respect individuals' silences and left unstructured time so that people could depart before group discussions. In this way, people could decline participation without being singled out by the community in socially damaging ways.

Assessing voluntary involvement is essential to produce valid research. Orphan children participating in group discussions as part of another child-protection study in Liberia all signed consent forms, yet we later realized that the children had been pressured to participate and told what to say by the administrators of the orphanages.

The study of mental health across cultures requires creativity, tenacity and time. But these investments must be made. There is a dearth of cross-cultural mental-health research to inform practice. Ultimately, collaborative partnerships make for better research and boost the utility of findings. To relieve suffering, we must learn what works best in each setting, and we must ethically build the capacity to do so. ■

Mónica Ruiz-Casares *is assistant professor in the Division of Social and Transcultural Psychiatry at McGill University in Montreal, Canada.*
e-mail: monica.ruizcasares@mcgill.ca

1. Barata, P. C. *et al. Soc. Sci. Med.* **62,** 479–490 (2006).
2. Ruiz-Casares, M. *Transcult. Psychiatry* http://dx.doi.org/10.1177/1363461514527491 (2014).
3. Ruiz-Casares, M. & Thompson, J. *Child. Geograph.* (in the press).
4. Osamor, P. E. & Kass, N. *Dev. World Bioeth.* **12,** 87–95 (2012).
5. Vreeman, R. *et al. BMC Med. Ethics* **13,** 23 (2012).

WALKER AND WALKER/GETTY

Strength in diversity

Richard B. Freeman and **Wei Huang** reflect on a link between a team's ethnic mix and highly cited papers.

Sticking with co-authors with similar surnames to yours might dent the impact of your work. The reason is unclear, but bibliometrics suggest that teams with greater ethnic diversity generate papers that make more of a splash in the scientific literature.

We analysed[1] 2.5 million research papers in which all of the authors had US addresses. Our study showed that US-based authors with English surnames were more likely have co-authors with English surnames than would occur by chance; those with Chinese names were more likely to have co-authors with Chinese names, and so on. The trend held for seven other groups, including Russian and Korean populations, between 1985 and 2008 in 11 scientific fields, including biomedicine, physics and geosciences.

The results hint that scientific research is much like the rest of social life. Studies of social networks find that people eat with, work with and generally connect with others similar to themselves, a tendency that some sociologists call homophily.

To the extent that surnames can be a proxy for ethnicity, homophily in scientific collaborations also seems to be related to a work's reception in the scientific community. After controlling for numbers of authors and for factors such as an ethnic groups' population density, we find that greater ethnic homogeneity among authors is associated with a paper's publication in lower-impact journals. It also predicts fewer citations. Papers with four or five authors of multiple ethnicities have, on average, one to two more citations than those written by authors all of the same ethnicity. This effect represents a 5–10% difference in the mean number of citations for a given publication.

What might explain this observation? Scientists with lacklustre or fewer papers may have a narrower pool of potential collaborators. Homophily is greater for authors with weaker publication records. But even when we compare work from authors with similar publication histories, homophily is still associated with lower-impact papers.

NETWORK EFFECTS

Teasing out the implications of these findings is difficult. Teams with members from diverse ethnic backgrounds may benefit from a greater variety of perspectives. Researchers have been shown to think differently when they work in diverse groups because they expect greater challenges to their ideas, or because small group dynamics are altered[2]. Given that communication can be hampered by linguistic or cultural differences[3], perhaps researchers make an extra effort to work through these challenges on research questions that are likely to have particular impact.

Network effects offer a different sort of explanation. A paper generated by a more diverse research group could tap into different networks and thus attract greater attention and citations, an effect observed in patents studies[4], and in inter-institution and international collaborations[5]. And although using journal impact factors to infer the quality of individual papers is controversial, that relationship, too, indicates that diverse teams publish stronger papers.

In other words, greater diversity of authorship might boost either the quality of the paper or the number of people who notice it, or both. One way to distinguish between the two would be to examine the terms, techniques and references in papers. If ethnic diversity contributes to productivity by widening ideas, papers from more-diverse collaborations should contain a wider range of scientific terms, use more varied equipment, procedures, or data and reference a wider range of previous work than papers from homogenous groups. In the biomedical sciences, the medical subject heading, or MeSH, terms would provide a natural measure, as might automated text analysis.

Another approach would be to model the probable impact of network effects on citations and then estimate the effect of co-authors in differently sized ethnic networks. This type of analysis could also be used to determine the mechanism by which inter-institutional or international collaborations often have greater impact than collaborations written at a single address.

Finally, we are studying the ethnic mix of collaborators who met at scientific meetings, and the impact of resulting papers. This would control for variation in the opportunity to meet people of different ethnicity, and could isolate people's preference for homophily or diversity.

These are questions worth pursuing. We need to work out what makes the most creative and effective scientific teams. ■

Richard B. Freeman *is director of the Science and Engineering Workforce Project at the National Bureau of Economic Research, and professor of economics at Harvard University in Cambridge, Massachusetts, USA.* **Wei Huang** *is a PhD candidate in economics at Harvard University.*
email: freeman@nber.org

1. Freeman, R. B. & Huang, W. *J. Labor Econ.* (in the press).
2. Apfelbaum, E. P., Phillips, K. W. & Richeson, J. A. *Perspect. Psychol. Sci.* **9,** 235–244 (2014).
3. Samovar, L. A., Porter, R. E., McDaniel, E. R. & Roy, C. S. *Communication Between Cultures* (Cengage Learning, 2009).
4. Kerr, W. *Rev. Econ. Stat.* **90,** 518–537 (2008).
5. Adams, J. *Nature* **497,** 557–560 (2013).

AUTUMN BOOKS

GENETICS

Under the skin

Nathaniel Comfort wonders at the enduring trend of misrepresenting race.

Is race biologically real? A clutch of books published this year argue the question. All miss the point.

Michael Yudell's *Race Unmasked* and Robert Sussman's *The Myth of Race* can be read as inadvertent retorts to former *New York Times* journalist Nicholas Wade's *A Troublesome Inheritance*, published while the former were in the press. Wade's book is by far the most insidious, but all three are polemics that become mired in proving (in Wade's case) or disproving (in the others') whether race is biological and therefore 'real'. This question is a dead end, a distraction from what is really at stake in this debate: human social equality.

A Troublesome Inheritance: Genes, Race and Human History
NICHOLAS WADE
Penguin: 2014.

Race Unmasked: Biology and Race in the 20th Century
MICHAEL YUDELL
Columbia University Press: 2014.

The Myth of Race: The Troubling Persistence of an Unscientific Idea
ROBERT WALD SUSSMAN
Harvard University Press: 2014.

Race is certainly real — ask any African American. It originated long before the science of genetics, as sets of phenotypes and stereotypes. These correlate with haplotypes, clusters of genetic variation. In this sense, race is genetically 'real'. But those correlations depend on judgement calls. Wade cites population-genetics studies that identify three principal races: caucasian, African and East Asian. Elsewhere he cites five, adding Australasian and Native American; or seven, splitting caucasians into people from Europe, the Middle East and the Indian subcontinent.

ILLUSTRATIONS BY DARREN HOPES

A study in *Scientific Reports* this year identified 19 "ancestral components", including Mozabites, Kalash and Uygurs (D. Shriner *et al. Sci. Rep.* **4,** 6055; 2014). Palaeogeneticist Svante Pääbo and others have revealed the underlying human genetic variation to be a series of gradients. Whether and how one parses that variation depends on one's training, inclination and acculturation. So: race is real and race is genetic, but that does not mean that race is 'really' genetic.

The completion of the draft human-genome sequence in 2000 led some optimists to forecast the end of race (one of them, Craig Venter, wrote the foreword to Yudell's book), but use of the term in the biomedical literature has actually increased since then. For clinicians, race is a matter of pragmatism. Although each of us is genetically and epigenetically unique, our ancestry leaves footprints in our genomes. Consequently, clinicians use familiar racial categories such as 'black' or 'Ashkenazi Jewish' as crude markers of genotypes, in a step towards individualized medicine. For them, the reality of race is immaterial; diagnosis and treatment are what count (see page 301).

Debates over the genetic reality of race, then, are not mainly scientific, but social. They deploy the cultural authority of science — considered society's most objective way of understanding the world — as a fig leaf for positions motivated explicitly or implicitly by ideology. All three of these books argue that if the proof or disproof of race is scientific, it must be true. The author must be right. More importantly, his opponents must be wrong.

For Wade, science proves that race is genetic. Much like Richard Herrnstein and Charles Murray's *The Bell Curve* (Free Press, 1994), his book moves smoothly through seemingly reasonable arguments that humans are still evolving, to end up at the retrograde conclusion that Europeans have become the world's richest and most powerful people mainly because they are genetically the most open, curious, innovative and hard-working. Also like *The Bell Curve*, Wade's book draws heavily on a long tradition of what historians refer to as scientific racism, particularly research connected to the Pioneer Fund, chartered in 1937 in part to "support study and research into the problems of heredity and eugenics" and, as Sussman shows, still deeply involved in eugenic and racial research. Despite such transparently political sources, Wade insists that his argument is based on ideology-free science. On 8 August, 139 population geneticists — including several on whose work Wade based his arguments — signed a letter to *The New York Times* declaiming his use of their results. Now that those whose work he once categorized as "scientific" instead of "ideological" have come out against the book, Wade has denounced them, too, as being motivated by politics.

By contrast, the 'race realist' and 'human biodiversity' (HBD) groups are delighted with Wade's book. For example, Jared Taylor, editor of the HBD magazine *American Renaissance*, applauds Wade's argument that (in Taylor's paraphrase), "foreign aid is probably wasted because poor countries are not genetically prepared for the institutions necessary for wealth". Other pillars of the race-realist movement, such as the website Stormfront and the writer John Derbyshire, gave Wade's book glowing reviews.

DEBATES OVER THE GENETIC REALITY OF RACE ARE NOT MAINLY SCIENTIFIC BUT SOCIAL.

For both Yudell, a historian of public health, and Sussman, a cultural anthropologist, science proves that race is cultural. In making this case, both devote considerable space to eugenics, the science and social movement concerned with human hereditary improvement. The eugenics movement — particularly in the United States in the early twentieth century and in Nazi Germany — offers a cornucopia of evidence of scientific racism. But, in focusing on the US movement's most egregious leaders, such as Charles Davenport, Madison Grant and Henry Fairfield Osborn, both Yudell and Sussman over-simplify. Eugenics was about much more than just race. Recent scholarship has documented the pervasiveness and adaptability of the US and UK eugenic creed and the complicated ways it mingled with race, public health and feminism. Sussman and Yudell both, for example, discuss the efforts of the distinguished black leader W. E. B. Du Bois against white-supremacist eugenics in the early twentieth century. But both fail to mention his concern over black "dysgenics" and his eugenically inflected "talented tenth" campaign, which sought to identify the "best of this [black] race". Not all eugenics is racist, and most racism is not — or not principally — eugenic.

For both Yudell and Sussman, the antidote to eugenic hereditarianism was cultural anthropology, developed by Franz Boas and his students in the late nineteenth and early twentieth centuries. Boas coined the word culture in its modern sense, and became perhaps the greatest opponent of the biological concept of race. He and his students studied human societies through an entirely cultural definition of human difference. Boas found, for example, that cranial characteristics that had been claimed to be innately racial were the result of differences in nutrition and overall health. Sussman and Yudell insist that Boasian anthropology scientifically proved that race is not genetic.

Their arguments then diverge. Sussman becomes, if possible, more polemical, whereas Yudell grows slowly less so. The former returns to the history of scientific racism, providing a passionate account of the continuing influence of the Pioneer Fund. Yudell's late chapters, by contrast, trace the struggle to strip racism from race science.

Yudell offers a rich analysis of the statements on race by the United Nations Educational, Scientific and Cultural Organization (UNESCO) — a string of contentious multidisciplinary reports that sought to document scientific knowledge on race while denouncing racism, starting in 1950. The project's myriad authors split into two factions. One, led by Boas's student Ashley Montagu, wanted to call race a fiction, a product of culture. The other insisted that genetics showed that race was real. Theodosius Dobzhansky, a brilliant population geneticist, was intellectually invested in the genetic concept of race, yet morally invested in anti-racism. "Dobzhansky's paradox", in Yudell's phrase, was how to save biological race theory without sounding racist. He never did — and nor have we, Yudell concludes poignantly.

A full-throated, intellectually rigorous anti-racism must critically assess both biological and cultural evidence about race. It must acknowledge that no work on race science can be free of ideology — and, precisely for that reason, it must not place historical actors before a moral green screen showing an image of contemporary values. Rather, it must set the stage for each scene with meticulous, empathetic historical detail. Such work would allow the scientific study of 'racial superiority' — inherently grounded in subjectivity and bias — to fall on its own sword. ■

Nathaniel Comfort *is at the Institute of the History of Medicine at Johns Hopkins University in Baltimore, Maryland. His latest book is* The Science of Human Perfection. *e-mail: nccomfort@gmail.com*

PHYSICS

In thrall to uncertainty

A history of how quantum theory has permeated Western culture refreshes **Jim Baggott**.

Quantum theory is the most accurate and precise description of the molecular, atomic, sub-atomic and sub-nuclear realms ever devised. It is also utterly exasperating. To anyone tutored in the language and the logic of classical physics, it is mathematically challenging, maddeningly bizarre and breathtakingly beautiful. As charismatic US physicist Richard Feynman warned: "Nobody understands quantum mechanics."

Given its recondite nature, quantum weirdness has entered popular culture in remarkable ways. What might otherwise have been the preserve of dry academic texts and stuffy scientific conferences has become common currency. Who hasn't heard of Heisenberg's uncertainty principle or Schrödinger's cat? Quantum ideas of space, time and matter inspired British artist Anthony Gormley's vast, enigmatic steel sculpture *Quantum Cloud* near London's O2 arena. And UK-based dramatist Tom Stoppard's 1988 play *Hapgood* blends the duality of the double agent with a quantum duality in which matter and light are both waves and particles. Both of these works are cited in *The Quantum Moment*, in which philosopher Robert Crease and physicist Alfred Goldhaber explore quantum theory's enduring cultural impact.

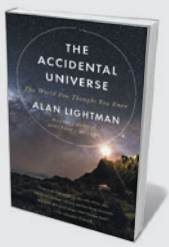
The Quantum Moment: How Planck, Bohr, Einstein, and Heisenberg Taught Us to Love Uncertainty
ROBERT P. CREASE AND ALFRED SCHARFF GOLDHABER
W. W. Norton: 2014.

Based on a class that the authors have taught for six years at Stony Brook University, New York, the book asks why quantum theory carries such a metaphorical punch — "wild and mysterious, packed with creative force" — and why it seems to be rediscovered in every generation. The authors' tale is structured along approximately linear historical lines, from Max Planck's discovery in 1900 that light can be described in terms of discrete 'bundles' of energy (quanta) to Bell's theorem, which opened the door from the 1960s onwards to some extraordinary experimental tests of the nature of our physical reality. Each chapter explores how some of the core ideas and principles that sprang from these historical moments have been absorbed into (inevitably mostly US) popular culture. This makes for an entertaining read, even for those already familiar with the story.

As Crease and Goldhaber explain, much of the impact on modern culture derives from what quantum theory has to say about the nature of reality. Arguably, centuries of observation, experimentation and strenuous intellectual endeavour were founded

NEW IN PAPERBACK

Highlights of this season's releases



The Accidental Universe: The World You Thought You Knew
Alan Lightman (Vintage, 2014)
Physicist and literary wizard Alan Lightman reflects on how our cosmos, potentially one among uncountable others, has fortuitously created the perfect conditions for life. He considers intricate symmetries in nature and the unfathomable vastness of space. This journey through seven overlapping 'universes' — frameworks for exploring recent research — culminates in a vision of humanity hooked on technology, gradually detaching itself from reality.

on scientists' expectation that the material world is composed of some kind of fundamental atoms of 'stuff'. Quantum theory, however, has rewarded these endeavours with phantom particles that, like waves, can be both here and there; a theoretical structure that tells us only what might happen (not what will); and quantum systems seemingly connected over great distances, giving rise to extended, non-local effects, or what Albert Einstein called "spooky action at a distance".

Einstein famously rejected the element of chance that lies at the heart of quantum theory, declaring that God does not play dice. He argued that quantum theory is not complete. Those scientists who, like Einstein, are less inclined to accept that we have reached an ultimate limit of what is knowable remain firmly in denial. So, in the past 40 years or so, the efforts of agents provocateurs such as John Bell and Tony Leggett have encouraged an orgy of sophisticated laser-based experiments to test the foundations of quantum physics — what I have elsewhere called "experimental philosophy". It is this work that has prompted the current interest in quantum cryptography, quantum computing and the teleportation of photons.

I have only one quibble with *The Quantum Moment*. Crease and Goldhaber support their narrative with 'interludes' after each chapter, designed to provide technical details, including some equations. The exposition here is a little drier than in the main chapters, but does not need to be. The material also necessarily repeats much of what has already been covered, which can become a little tedious. The authors suggest that readers might prefer to skip these interludes; for linear readers like me, that does not really work.

Those versed in quantum theory's practical applications might be tempted to dismiss its many manifestations in popular culture as what the authors call "fruitloopery". And certainly, there is a lot of nonsense out there. But, as Crease and Goldhaber make abundantly clear at several points, many esteemed physicists (who should probably know better) have been more than willing to indulge their inner metaphysician in research papers and popularizations on the mistaken principle that, as the Canadian philosopher Marshall McLuhan once put it, "mud sometimes gives the illusion of depth".

Thus we smile at the comical pronouncements on physics by US actress Shirley MacLaine, until the authors point out that she could be paraphrasing similar pronouncements made 55 years earlier by the British physicist James Jeans. I have elsewhere argued that contemporary theoretical physics has become rather self-indulgent and self-referential, a malaise that I have called fairy-tale physics. Deep questions about the nature of reality tend to provoke this kind of response, and it surely finds its origin in the quantum moment. ■

Jim Baggott *is the author of* The Quantum Story *and* Farewell to Reality. *He is based in Reading, UK.*
e-mail: jim@logosconsulting.co.uk

LINGUISTICS

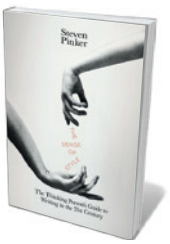
The write stuff

Steven Pinker's provocative treatise on language use and abuse would benefit from more data, finds **Paul Raeburn**.

No conversation about the science of language can get very far without a mention of Steven Pinker, the Harvard University cognitive scientist who has not yet made linguistics as popular as football — but is working on it. In *The Sense of Style*, he wants to give us the cognitive science, linguistics and psychology behind classic debates over proper English, from passive voice to split infinitives.

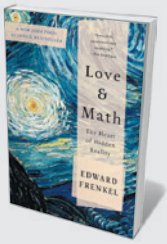
Plenty of others have given us stuffy decrees intended to end the interminable wrangling, but Pinker is different. He is unhappy with the classic style manuals — including revered texts such as *Strunk & White* (William Strunk and E. B. White's *The Elements of Style*) or *Fowler's Modern English Usage*. We need a new guide "infused by the spirit of scientific skepticism", he writes, using grammar and research on "the mental dynamics of reading" to replace edicts with evidence. Pinker gave us the science in *The Language Instinct* (William Morrow, 1994); in *The Sense of Style* he sets out to offer its practical application.

He covers much of the same ground as the classic guides, including frequently misused words ("fulsome" and "noisome") and the serial comma. His problem with *Strunk & White*, however, is that the authors lack tools for analysing language, and so end up "vainly appealing to the writer's 'ear'". That's on page two. By page three, he is challenging the manual's dismissal of the passive voice. Linguistic research, he later writes, has shown that the passive actually "allows the writer to direct the reader's gaze, like a cinematographer choosing the best camera angle". What research, exactly? Pinker does not tell us. His views are informed by psycholinguistics; that is his day job. But he promises us science, so I expected to see data. However, in this instance, and in many others, the data are not there.

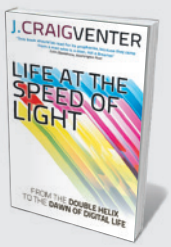


The Sense of Style: The Thinking Person's Guide to Writing in the 21st Century
STEVEN PINKER
Allen Lane: 2014.

Similarly, Pinker's view on infinitives is to split them "if you need to", a conclusion backed by dictionaries and style manuals — not research. And when he quotes with admiration the opening line of Richard Dawkins' *Unweaving the Rainbow* (Houghton Mifflin, 1998) — "We are going to die, and that ▶



Love and Math: The Heart of Hidden Reality
Edward Frenkel (Basic Books, 2014)
With infinite passion, media-feted professor Edward Frenkel shares his rise to mathematical greatness against a tide of Russian anti-Semitism. Appeasing maths-haters, he uses a borscht recipe to explain quantum duality. (See Marcus du Sautoy's review: *Nature* **502,** 36; 2013.)



Life at the Speed of Light
J. Craig Venter (Abacus, 2014)
Biologist J. Craig Venter shares his life's work of catalysing progress in biological engineering, sequencing the human genome and ultimately creating the first "synthetic cell" (*Mycoplasma mycoides* JCVI-syn1.0). (See Nathaniel Comfort's review: *Nature* **502,** 436–437; 2013.)

makes us the lucky ones" — he offers a detailed explanation of why it works that is, again, short on science.

Pinker is a good writer and a deeply humanistic one, and there are many bright moments here. His lists explaining right and wrong usage with a range of examples (enervate means to sap, not energize) are a useful desk reference. Among numerous good tips is one on, as Pinker has it, "the compulsion to name things with different words when they are mentioned multiple times". "Herons are herons," he writes, not "long-legged waders, azure airborne aviators, or sapphire sentinels of the sky".

At times, however, Pinker's own writing verges on the incomprehensible. Consider his critique of this sentence: "Toni Morrison's genius enables her to create novels that arise from and express the injustices African Americans have endured." Some might say 'her' is an error, because an adjective ('Toni Morrison's') cannot be the antecedent of a pronoun. But Pinker explains it this way: "*Toni Morrison's* is not an adjective, like *red* or *beautiful*; it's a noun phrase in genitive case. (How do we know? Because you can't use genitives in clear adjectival contexts like *That child seems Lisa's* or *Hand me the red and John's sweater*.)" After reading that several times, I think I know what he means. But it is tough to get through.

Pinker also reveals himself at the outset to be not a prescriptivist, like Strunk and White, but a descriptivist, who sees language as "a wiki that pools the contributions of millions of writers and speakers".

I agree: we make the language. But if that is the case, science probably can't do any better than *Strunk & White* at dictating style. The only legitimate data come from the people. So maybe it is too soon to jettison the classic style manuals: I suspect much of Pinker's sense of style comes less from his science than from his own wonderful writer's ear. ■

Paul Raeburn *is the author of four books and the chief media critic for the Massachusetts Institute of Technology's Knight Science Journalism Tracker.*
e-mail: paulraeburn@nasw.org

EVOLUTION

Tribes like us

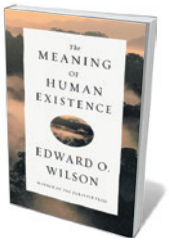
Tim Lenton is intrigued by E. O. Wilson's sweeping perspective on humanity's past — and possible futures.

What of that ultimate existential question, the meaning of life? Aristotle saw it as the achievement of happiness. UK comedy troupe Monty Python suggested that it involves reading "a good book every now and then". In *The Meaning of Human Existence*, biologist E. O. Wilson offers a good book that adds to such prescriptions, but readers seeking a sense of purpose will be disappointed. What Wilson is after is really a deeper understanding of human existence.

Still, there can be few better guides through our species' past journey and potential for the future. Wilson provides the literary equivalent of a greatest-hits album, giving us a pithy synthesis of his formidable body of work from *Sociobiology* (Harvard University Press, 1975) to *The Social Conquest of Earth* (Liveright, 2012), with a liberal dose of *Consilience* (Little, Brown, 1998). The result is a provocative and beautifully written collection of essays, although one that struggles to be more than the sum of its parts.

In the opening section, Wilson introduces his central premise that humans, like his beloved ants, are eusocial animals. Some individuals reduce their own lifetime reproductive potential so that they can raise the offspring of others (think of grandmothers after menopause). Key to the origin of eusociality is the creation of a nest, from which some of the population undertake risky foraging while the remainder stay safe at home. Wilson argues that our unique intelligence began to evolve when our ancestors tamed fire to cook, settled around the campsite and sent a fraction of the group off to risk life and limb hunting down energy-rich meat.

Thus began a tension between acting for ourselves and acting for our group, which Wilson argues is at the heart of our conflicted human nature. Here he parts company with most evolutionary theorists, revisiting an already acrimonious debate (aired in *Nature*) over the origin of eusocial traits. Wilson originally supported evolutionary biologist W. D. Hamilton's theory of inclusive fitness, in which the costs of altruism can be rationalized if they are outweighed by the product of the benefits to recipients and the recipients' relatedness to the altruist. But in 2010, he and some colleagues rejected it (M. A. Nowak *et al. Nature* **466,** 1057–1062; 2010). In its place, they argued for a mixture of individual and group-level selection.

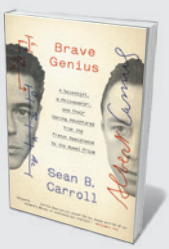


The Meaning of Human Existence
EDWARD O. WILSON
Liveright: 2014.

Back from the firmly prodded ants' nest of evolutionary theorists came a predictably forceful defence (see, for example, P. Abbot *et al. Nature* **471,** E1–E4; 2011), but Wilson remains unmoved by this stinging riposte. The frustration for the neutral reader is that both sides agree that the gene is the fundamental unit of selection, so the squabble is over different flavours of standard evolutionary theory. Neither side seems to see the Pythonesque irony of fighting over how to understand cooperation. Still, nothing could better demonstrate the tribal nature of humanity, which provides a focus for the rest of the book.

Wilson's enthusiasm for a mixture of individual and group-level selection goes further, as he struggles to resist an "oversimplistic" portrayal that "individual selection promoted sin, while group selection promoted virtue". The inconsistency in this

The Compatibility Gene
Daniel M. Davis (Penguin, 2014)
At the heart of our immunological-response systems lie 'compatibility genes', which determine each body's capacity to fight diseases or accept medication. Immunologist Daniel Davis explores these genes' roles in successful skin grafts, ill-fated pregnancies and more.



Brave Genius
Sean B. Carroll (Broadway, 2014)
Against the tumult of the Second World War, biologist Sean Carroll tells the interwoven stories of philosopher Albert Camus and geneticist Jaques Monod, friends who worked for the French resistance and won Nobel prizes. (See Jan Witkowski's review: *Nature* **501,** 487–488; 2013.)

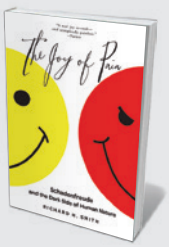
is soon exposed when he argues that religion has been crucial in reinforcing group-level tribalism, but is a collective sin that humanity needs to grow out of. One wonders what the publishers were thinking when they put on the dust jacket the promise of Wilson addressing "our greatest moral dilemma since God stayed the hand of Abraham", given that inside, he decries belief in God with Dawkinsian fervour.

Yet Wilson's route to species self-knowledge is rather omniscient, because it involves comparing ourselves to other known or imagined life forms, be they ants or aliens. As we decimate biodiversity, leaving ourselves lost in an Age of Loneliness — the 'Eremocene' — Wilson looks skyward for salvation. He is excited about exoplanets and brimming with existential confidence that we are not alone in the Universe, offering a very anthropomorphic "portrait of E.T.".

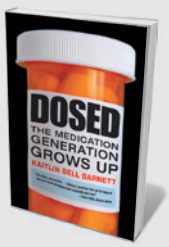
Wilson's imaginary aliens are, I think, really his prescription for humanity's future. They have chosen not to supplement their intelligence or engineer their genetics, because their technological creations have long surpassed them physically and intellectually. They are not so foolish as to attempt interstellar travel — were it possible — because they have worked out that invading an independently evolved world would be a biological train wreck. Instead, they have made peace with their home planet and achieved a long-term sustainable state.

Back in the here and now on Earth, Wilson argues that we should relish our evolutionary legacy of internal conflict and the creativity it sparks. He sees a short future for scientific progress but a long one for the humanities, arts and social sciences. Surprisingly optimistic that brain-activity mapping is going to solve the riddle of human consciousness sooner rather than later, Wilson feels that we will be left clutching the sensation of free will, which he thinks is just an adaptation necessary for our sanity. If the resulting nihilism does not lead us to despair, the way forward will be to unify the sciences and humanities to reach a higher state of human "meaning". Anyone fancy the ride? ■

Tim Lenton *is professor of Earth system science at the University of Exeter, UK, and co-author with Andrew Watson of* Revolutions That Made the Earth.
e-mail: t.m.lenton@exeter.ac.uk



The Joy of Pain
Richard H. Smith (Oxford Univ. Press, 2014)
Psychologist Richard Smith explores the roots of *Schadenfreude* (joy in others' pain) in society, from reality television thriving on public humiliation to cases of envy-incited crimes, including Nazi persecution of Jewish people. (See Dan Jones' review: *Nature* **500,** 147; 2013.)



Dosed: The Medication Generation Grows Up
Kaitlin Bell Barnett (Beacon, 2014)
Journalist Kaitlin Bell Barnett, herself medicated in youth, tells the stories of five people who from childhood have been treated with psychotropic drugs for conditions such as depression and attention-deficit hyperactivity disorder, addressing their sense of lost freedom and identity.

CLIMATE POLICY

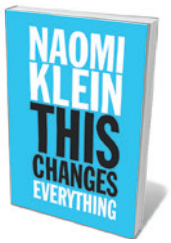
A societal sea change

Nico Stehr ponders Naomi Klein's call for strategic mass action on climate change.

This year, New Zealand became the first country to acknowledge climate change when granting residence on humanitarian grounds — in this case, to a family from the low-lying Pacific island nation of Tuvalu. That the environmental and human realities of climate change are tangled up in such legal, economic and political complexities is the focus of social activist Naomi Klein's *This Changes Everything*.

Klein's journey to this realization began in 2009 when, in the middle of the international economic meltdown, she first recognized the magnitude of climate change. As she states: "I denied climate change for longer than I care to admit." Five years on, she has synthesized her thinking about the dual financial and environmental disasters, arguing that the "market fundamentalism" favoured by the United States and the United Kingdom — involving privatization of public services, deregulation of corporate activities and reduced public spending — has "systematically sabotaged our collective response to climate change". Klein now advocates a mediating social force between climate science, politics and individual responsiveness and responsibility — in essence, "mass social movements" aiming to reduce fossil-fuel use and push for adaptation measures and behavioural change.

Klein's book — a combination of polemic, manifesto and analysis — covers much familiar territory. We get the history of global-warming discussions, the phenomenon of climate denialism and today's economic order, including the clash of interests between trade regimes and climate policy. As Klein notes, the "liberation of world markets, a process powered by the liberation of unprecedented amounts of fossil fuels", is now helping to accelerate the melting of Arctic ice. She also discusses responses to the climate emergency, including geoengineering schemes, as well as resistance to large-scale



This Changes Everything: Capitalism vs the Climate
NAOMI KLEIN
Simon and Schuster/ Allen Lane: 2014.

mining projects in various parts of the world. She includes her own eye-witness reportage from the front line, such as Canada's oil sands and the 2010 Deepwater Horizon oil spill in the Gulf of Mexico.

Klein acknowledges that extraction and use of fossil fuels is hardly analogous to social or political oppression such as slave ownership or gender discrimination, but argues that historical movements against such practices show how societal pressure can build, pushing governments to act. Mass action focused on climate change could have a transformative effect on societies, she posits, empowering the poor to demand rights and services.

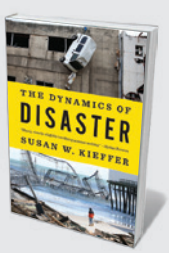
In arguing that the "fundamentalist" changes to the structure of capitalism have stymied such a transformation, Klein overstates the case, however. Although privatization and deregulation triumphed in the 1980s in the United States and the United Kingdom, they did not in significant parts of the rest of the world, as she claims. And whether the US economic order exemplifies a secular trend towards global dominance remains an open issue. The conditions for change could be more favourable than Klein thinks, especially when it comes to the removal of ideological roadblocks to improving the ethics of markets on the basis of moral rather than mere monetary motives of production and consumption.

Of course, many other impediments to governing climate change remain. Klein is too optimistic in claiming that the immediately relevant solutions, such as adaptation or reducing fossil-fuel use, have broadly been long understood. The governance of climate change, not merely mitigation and adaptation, is a tortuous problem and is hardly well developed theoretically, but it remains a key area of interdisciplinary research and real-world policy. We are just at the beginning of this difficult intellectual and practical journey. Klein recognizes the Sisyphean tasks ahead and proposes economic, legal and social measures that would enable better governance, such as the reform of trade law and the prevention of fossil-fuel extraction through the recognition of indigenous peoples' rights to oil- and coal-rich land.

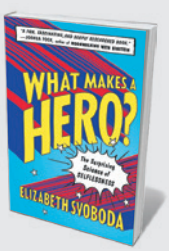
The special appeal of Klein's position is her insight that any successful effort to curb emissions or adapt to climate change demands popular, pragmatic and sensible transformative goals that go well beyond mere fencing in. In contrast to climate scientists and observers such as James Hansen and James Lovelock, she is not an advocate of "inconvenient democracy" — that is, the claim that certain forms of democratic governance need to be overcome before climate change can be tackled effectively.

Whether *This Changes Everything* has identified the potential catalyst that will bring about an alternative future remains uncertain. Still-to-be-born large social movements around the world could act as a countervailing force to 'business as usual' and, as Klein puts it, "simultaneously clear some alternative pathways" to safer futures for humankind. Klein has explored early manifestations of such resistance among smaller, often ad hoc local social groups around the world. But her message is still delivered with a strong North American accent. ■

Nico Stehr *is Karl Mannheim Professor of Cultural Studies at Zeppelin University in Friedrichshafen, Germany.*
e-mail: nico.stehr@t-online.de



The Dynamics of Disaster
Susan W. Kieffer (W. W. Norton, 2014)
Geologist Susan Kieffer showcases Earth's most destructive processes, highlighting geographical discrepancies in disaster preparedness. In 2010, for example, similar-energy earthquakes caused over 50,000 deaths in Haiti, yet none in New Zealand. (See Roger Bilham's review: *Nature* **502,** 438–439; 2013.)



What Makes a Hero?: The Surprising Science of Selflessness
Elizabeth Svoboda (Current, 2014)
Would you risk your life for a stranger's? Survival instinct would suggest not, but science writer Elizabeth Svoboda finds that heroism comes naturally to some, and others can learn altruism using methods such as compassion meditation.

INTERNET

Technology and its discontents

Jaron Lanier surveys four studies probing the vexed nexus of mind and digisphere.

Digital technology is remaking the cognitive environment in which human brains develop and function. This swift revolution is inevitably sparking much hard thinking. Books by neuroscientists Susan Greenfield and Daniel Levitin, and writers Nicholas Carr and Paul Roberts, propose either adaptation to the changes — self-help strategies to compensate for emerging cognitive misalignments — or critiques of the overall transformation.

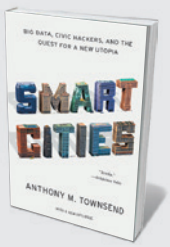
Greenfield's *Mind Change* takes the latter approach. It proposes that global climate change can serve as a useful metaphor for how human minds — our inner environments — are, in her view, being recklessly altered by digital technologies. Greenfield argues that because the human brain is remarkably plastic in youth, it is not unreasonable to ask how recently introduced, ubiquitous digital designs (such as those of social networks or reading tablets) might affect brain development. The acquisition of speech and reading can affect human brain architecture, but there has been little precedent for the kind of sudden, uniform, pervasive change in children's cognitive environments posed by these

Mind Change: How Digital Technologies Are Leaving Their Mark on Our Brains
SUSAN GREENFIELD
Rider: 2014.

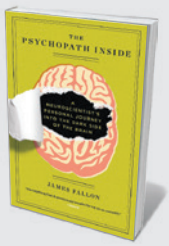
The Organized Mind: Thinking Straight in the Age of Information Overload
DANIEL J. LEVITIN
Dutton: 2014.

The Impulse Society: What's Wrong With Getting What We Want?
PAUL ROBERTS
Bloomsbury: 2014.

The Glass Cage: Automation and Us
NICHOLAS CARR
W. W. Norton: 2014.



Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia
Anthony M. Townsend (W. W. Norton, 2014)
As technology infiltrates urban life, Anthony Townsend observes how cities evolve in the digital sphere, from parking apps in Germany to crowd-sourced maps of African slums. (See Melanie Moses' review: *Nature* **502,** 299–300; 2013.)



The Psychopath Inside
James Fallon (Current, 2014)
After confusing his own brain scan with a psychopath's, neuroscientist James Fallon trawled his past and genealogy. Assembling evidence from obsessive–compulsive disorder to violence in his family history, Fallon considers how nurture may overcome nature.

technologies. How might they affect the sense of identity or organic memory, for instance? Although she sometimes seems to push her argument beyond the reach of current research, Greenfield asks key questions — such as whether the next generation will think less critically than their forebears. And she broadly outlines the kind of research and policy agenda needed to address such haunting unknowns.

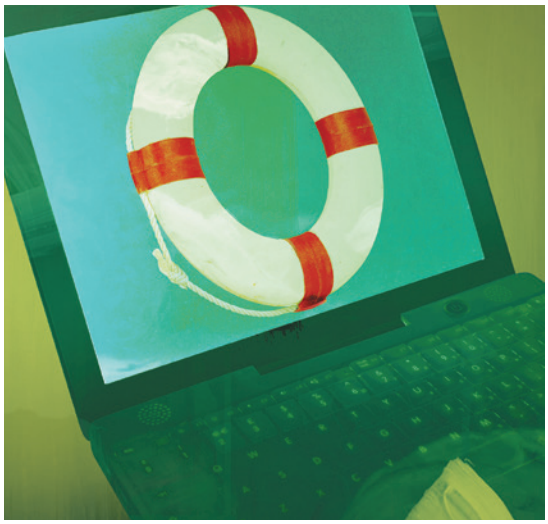
She occasionally veers into alarmism, for instance when discussing speculative links between the apparent rise in autism and the rise in the use of particular digital environments. However, some of Greenfield's caution may be justified. The neuroscience and cognitive-science communities that overlap with digital-technology developments often rely on the technology industry for support or cooperation, so it is especially important that they are not swayed by that industry's extreme enthusiasms. For all its faults, *Mind Change* is an important presentation of an uncomfortable minority position. It should be read by technologists in particular, as a check on self-congratulation.

By contrast, in *The Organized Mind*, Levitin takes the self-help approach. Accepting the design of information technology and today's information deluge as givens, he explores better brain function in that context. Our networked age often confounds the human mind, he notes, because of the kinds of cognitive quirks investigated by psychologists Daniel Kahneman and his late colleague Amos Tversky — notably Kahneman's idea of two brain systems, one 'quick and dirty' and the other slower and more reasoned. Levitin's strategy for overcoming such quirks is a set of tricks. To bypass poor intuitions about statistics, for instance, he suggests assessing data using a simple four-fold diagram.

Levitin's presentation is sensible and actionable, but I suspect that his audience is the sub-population residing between the extremes of technical ability. This group holds much of society's money and power: our highly technical society is for the most part guided by semi-technical people.

Carr's *The Glass Cage* — a meditation on automation, from apps-for-everything to self-driving cars — asks at the start how we should define a human being in such an era. Does automation change the sense of how people act, learn, or find value in their lives and each other? Carr tells contemporary and historical tales of technologists and entrepreneurs dripping with hubris, such as aviation wizard Wilbur Wright, and of people struggling with a sense that they are becoming denatured by a reliance on automation.

Carr can be understood as part of a literary movement that does not reject technologies. Rather, it rejects ceding what Carr calls "choices about the texture of our daily lives" to technologists and their businesses. That stance is a tightrope walk: one must move forward, succumbing neither to Luddite tendencies nor to the seductions of hot technological trends.



AUTOMATION IN THE AGE OF CLOUD COMPUTING IS OFTEN A FAKE FRONT. REAL PEOPLE, ANONYMIZED AND DEVALUED, ARE THE SOURCES OF THE 'BIG DATA'.

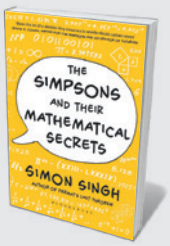
Carr is one of our most accomplished tightrope walkers. However, *The Glass Cage* does fall prey to a flawed conceit. Automation in the age of cloud computing is often a fake front. It is real people, anonymized and unvalued, who are the sources of the 'big data' that allow cloud algorithms to function. Automatic language translation is made possible only through daily sampling of human translators' work. Celebrating how people are contributing to technology in new ways could address some of the problems Carr decries, whether economic or cognitive.

For *The Impulse Society*, Roberts draws on the work of research psychologists such as Walter Mischel, who has studied delayed gratification. More lament than prescription, the book considers the many ways in which technologies encourage an infantile desire for immediate gratification. What is most striking about Robert's critique is its panoramic sweep. During the financial crises of the past decade, for instance, an urge for an instant 'hit' cropped up among individual borrowers keen on home ownership, lenders set on unbelievable deals, and shareholders eager for soaring security valuations. At every level people were disabled by a common infatuation with false gold proffered by digital networks.

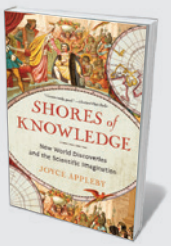
Roberts trips a bit towards the end of his book: he calls for a resurgence of traditional community as an alternative to the modern trend towards impatience. The book's ultimate programme seems sentimental and ill-matched to the theatre in which the troubles arise.

Taken together, these four books reveal a frontier of human experience. We are rapidly changing society, and in the course of it potentially laying our brains open to change. We must now become both competent and wise in our powers — not simply resisting or embracing new media technologies, but becoming instead more self-aware and discerning in relation to them. ■

Jaron Lanier *is a computer scientist with Microsoft Research. His latest book is* Who Owns the Future?
e-mail: jalani@microsoft.com



The Simpsons and their Mathematical Secrets
Simon Singh (Bloomsbury, 2014)
US television series *The Simpsons* is craftily dotted with maths jokes by numerate writers who chose comedy over academia. Physicist Simon Singh exposes and explains gags of varying complexity, although all can chuckle at Homer's naive belief in an "infinity plus one".



Shores of Knowledge: New World Discoveries and the Scientific Imagination
Joyce Appleby (W. W. Norton, 2014)
Six centuries of overseas exploration is lucidly charted by historian Joyce Appleby. While voyagers exulted over exotic species, the spread of disease to indigenous peoples exposed the high price of scientific discovery. **Emily Banham**

Correspondence

Ebola virus control needs local buy-in

International guidelines describe effective measures for the prevention and control of the Ebola virus. But we need more practical information on how to implement these measures, including potential therapies and a safe vaccine, in non-Western settings (*Nature* **513,** 13–14; 2014).

Culturally tailored procedures must be detailed in international public-health protocols, drawing on the expertise of medical anthropologists. Such measures will also help to overcome mistrust of authorities and field workers among affected populations.

The current Ebola outbreak will not be controlled without the local population's understanding and cooperation.

Gilles Guerrier *Hôtel-Dieu and Cochin Hospitals, Paris Descartes University, France.*
Eric D'Ortenzio *Solthis, France; and Bichat Hospital, Paris, France.*
guerriergilles@gmail.com

Climate models: sunk by humans?

Paul Palmer and Matthew Smith's argument that human adaptation to climate change should be incorporated into climate-projection models is entirely reasonable (*Nature* **512,** 365–366; 2014). However, I suspect that doing so could render such models essentially useless.

Climate models are created with the intention of providing predictions that are more reliable, and as such must always wrestle with the bias–variance dilemma. Introducing human responses to climate change will make this issue much more challenging than it already is — and perhaps hopelessly so.

To put it bluntly, one does not need to be an expert in modelling non-linear systems to recognize that the best answer to the question 'How can we get more precise predictions?' is never 'Add lots more variables'.

Robert A. J. Matthews *Didcot, UK.*
rajm@physics.org

Climate models: use archaeology record

Archaeologists and historians have long investigated societal responses to climate change (see P. Palmer and M. Smith *Nature* **512,** 365–366; 2014). These records are an underused resource in current climate-adaptation research, but offer scope for highly integrative meta-analyses that would be useful to climate scientists, science advisers and policy-makers, and could provide important information for local outreach efforts.

Risk-reduction researchers have pointed out that responses to climate change are a mix of contemporary industrial (technological) measures and pre-industrial (social and community-based) ones. However, the use of palaeoenvironmental data by the Intergovernmental Panel on Climate Change as a basis for drawing up future climate-change scenarios is not matched by an equally sophisticated use of 'palaeosocietal' data for investigating human impacts and adaptive pathways.

Archaeological and historical data could provide a solid evidence base for effective adaptations to climate change. Expanding the chronological scope of climate-adaptation research into deep time would vastly enlarge the database of available case studies without getting into the tricky issues of data access and legal sensitivity. In effect, this approach draws on natural experiments in history to learn from the past (see R. Van der Noort *Climate Change Archaeology* Oxford Univ. Press; 2013).

Felix Riede *Aarhus University, Højbjerg, Denmark.*
f.riede@cas.au.dk

Monitor Brazil's fish sampling closely

Brazil's aquaculture and fisheries secretary decreed last month that 2,000 different species of ornamental fish can be legally removed from the Brazilian Amazon. The fish will be farmed to supply the aquarium trade. This raises concerns about over-exploitation and threats to biodiversity, particularly given the poor record of inspection and reinforcement by the country's environmental agencies (see A. L. B. Magalhães and J. R. S. Vitule *Science* **341,** 457; 2013).

The new ruling could stimulate indiscriminate extraction, biopiracy, fish trafficking and the escape of farmed species into ecoregions of the country where they are not native. Close monitoring must be a priority.

We should be educating people about how to conserve Brazil's exuberant aquatic diversity, not encouraging its plunder.

Jean R. S. Vitule *Federal University of Paraná, Curitiba, Brazil.*
Flávia D. F. Sampaio *Federal Institute of Paraná, Curitiba, Brazil.*
André L. B. Magalhães *Pontifical Catholic University of Minas Gerais, Belo Horizonte, Brazil.*
andrebiomagalhaes@gmail.com

Shale gas is a fraught solution to emissions

Qiang Wang suggests that shale gas might be used as a bridging fuel to cap China's carbon emissions (*Nature* **512,** 115; 2014). Extraction and development problems could make this difficult.

The greenhouse-gas footprint of shale gas is much bigger than that of coal. Shale gas emits less carbon dioxide than coal or oil when burnt, but the methane produced during the extraction process has a global-warming potential 70 times that of CO_2 (see R. W. Howarth *et al. Nature* **477,** 271–275; 2011).

Following the US shale-gas boom, China devised a plan to extract its own gas resources. This proved difficult and expensive owing to limited water availability and because the gas is located at depth under large amounts of subsurface clay.

Furthermore, extraction might compromise the country's already stressed aquatic environments (H. Yang *et al. Nature* **499,** 154; 2013) and increase seismic activity — important factors in densely populated areas such as southwest China.

As a result, Chinese shale-gas production in 2013 was only around 3% of that planned for 2015. Last month, this forced the country to halve its production target for 2020 (see go.nature.com/h5mtza; in Chinese).

In our view, China would be better off investing more in renewable energy and improving energy efficiency.

Hong Yang *University of Oslo, Norway.*
Julian R. Thompson *University College London, UK.*
hongyanghy@gmail.com

Aristotle's suspect statistical skills

In his review of Armand Marie Leroi's book *The Lagoon* (*Nature* **512,** 250–251; 2014), Roberto Lo Presti rightly praises Aristotle's observational skills. But the philosopher may not have been as adroit in numeracy as he was in biology.

Aristotle famously declared that "males have more teeth than females in the case of men, sheep, goats, and swine" (see his *History of Animals*, Book 2, Part 3). This was an obvious sampling mistake, which bears out the importance today of a strong statistical foundation for biological curricula.

Taner Z. Sen *Ames, Iowa, USA.*
tanerzsen@gmail.com

GATM gene variants and statin myopathy risk

ARISING FROM L. M. Mangravite *et al*. *Nature* **502,** 377–380 (2013); doi:10.1038/nature12508

Mangravite *et al.*[1] identified six expression quantitative loci (eQTLs) that interacted with simvastatin exposure by using 480 lymphoblastoid cell lines exposed to β-hydroxy simvastatin acid *in vitro*. One of these SNPs (rs9806699) within the glycine amidinotransferase (*GATM*) gene was shown to have an association with statin-induced myopathy in two independent cohorts ($n = 172$ myopathy cases), conferring a protective effect (odds ratio = 0.61, 95% confidence interval = 0.39–0.95, $P = 0.03$). Our genotyping results from statin myopathy patients do not appear to replicate this finding. There is a Reply to this Brief Communication Arising by Mangravite, L. M. *et al. Nature* **513,** http://dx.doi.org/10.1038/nature13630 (2014).

Using primarily the UK Clinical Practice Research Datalink, an electronic healthcare record database, we recruited 145 cases with statin-induced myopathy and 537 statin-exposed control patients[2]. In addition, five patients meeting our case inclusion criteria were identified prospectively through a tertiary adult muscle clinic. Our myopathy phenotype was defined as serum creatine kinase levels of greater than 4 × upper limit of normal (ULN) or clinical record of rhabdomyolysis concurrent with statin prescription. In a proof-of-concept study, using a subset of patients (78 cases, 372 controls)[3] we were able to show an association between the *SLCO1B1*5* allele (rs4149056) and both statin-induced myopathy (odds ratio = 2.1, 95% confidence interval = 1.3–3.1) and severe myopathy ($n = 23$, odds ratio = 4.1, 95% confidence interval = 2.1–8.2), consistent with the genome-wide association study (GWAS) findings from the SEARCH collaborative[4].

We have undertaken genotyping for the rs9806699 *GATM* single-nucleotide polymorphism (SNP) in our cases and drug-exposed control patients ($n = 150$ and 587, respectively, after quality control) in order to attempt replication of the association shown by Mangravite *et al.*[1]. However, we were unable to show a significant difference in the minor allele frequency of rs9806699 between myopathy cases (MAF = 0.28) and controls (MAF = 0.30) (odds ratio = 0.94, $P = 0.68$). The MAF in our cases was similar to that identified in controls in the paper by Mangravite *et al.*[1]. By limiting cases to just those with 'severe' myopathy (creatine kinase > 10 × ULN or rhabdomyolysis) ($n = 37$), we again failed to show a significant difference in MAF between cases and controls (odds ratio = 0.94, $P = 0.83$). Further analysis restricted to patients only receiving simvastatin (99 cases, 344 controls) also did not demonstrate an association between rs9806699 and risk of either myopathy (odds ratio = 1.12, $P = 0.49$) or severe myopathy ($n = 26$, odds ratio = 1.42, $P = 0.24$). Analysis restricted to the 120 cases that were not on drugs known to interact with statins also did not change the result. We have also undertaken genome wide analysis of 128 myopathy cases (Illumina Human OmniExpress Exome 8v1), and comparison with the WTCCC2 (Wellcome Trust Case Control Consortium 2) genotype data also did not show any association between statin myopathy (generalized or severe) and any of 90 typed or imputed SNPs within the *GATM* gene locus.

In conclusion, we have not been able to replicate the association between the rs9806699 *GATM* SNP and statin myopathy reported by Mangravite *et al.*[1] in an independent sample set despite the fact that all patients were of European ancestry and had similar statin-myopathy phenotypes. This association will need to be assessed in more patients, and through an individual patient-data meta-analysis to determine, first, whether the SNP is relevant to a sub-phenotype of statin myopathy, and second, its clinical and mechanistic relevance.

D. F. Carr[1], A. Alfirevic[1], R. Johnson[1], H. Chinoy[2], T. van Staa[3] & M. Pirmohamed[1]

[1]Wolfson Centre for Personalised Medicine, Department of Molecular and Clinical Pharmacology, Institute of Translational Medicine, University of Liverpool, 1–5 Brownlow Street, Liverpool L69 3GL, UK.
email: Ana.Alfirevic@liv.ac.uk

[2]NIHR Manchester Musculoskeletal Biomedical Research Unit, Institute of Inflammation and Repair, University of Manchester, Oxford Road, Manchester M13 9PT, UK.

[3]Health eResearch Centre, Farr Institute for Health Informatics Research, University of Manchester, 1.311 Jean McFarlane Building, Oxford Road, Manchester M13 9PL, UK.

Received 19 December 2013; accepted 16 June 2014.

1. Mangravite, L. M. *et al*. A statin-dependent QTL for *GATM* expression is associated with statin-induced myopathy. *Nature* **502,** 377–380 (2013).
2. O'Meara, H. *et al.* Electronic health records for biological sample collection: feasibility study of statin-induced myopathy using the clinical practice research datalink. *Br. J. Clin. Pharmacol.* **77,** 831–838 (2014).
3. Carr, D. F. *et al. SLCO1B1* genetic variant associated with statin-induced myopathy: a proof-of-concept study using the clinical practice research datalink. *Clin. Pharmacol. Ther.* **94,** 695–701 (2013).
4. The SEARCH Collaborative Group. *SLCO1B1* variants and statin-induced myopathy—a genomewide study. *N. Engl. J. Med.* **359,** 789–799 (2008).

doi:10.1038/nature13628

GATM locus does not replicate in rhabdomyolysis study

ARISING FROM L. M. Mangravite *et al*. *Nature* **502,** 377–380 (2013); doi:10.1038/nature12508

All HMG-CoA reductase inhibitors (statins) can cause muscle injury ranging from asymptomatic elevations in creatine kinase levels to severe muscle breakdown (rhabdomyolysis) leading to kidney failure and death[1], and the genetic variants responsible for this uncommon adverse drug reaction remain largely undiscovered. Mangravite *et al.* reported a new locus in the gene *GATM* (rs9806699) that was associated with a decreased risk of muscle injury in two case-control studies of myopathy (odds ratio, 0.60)[2]. In a larger case-control study of statin-related rhabdomyolysis, a more severe form of muscle injury, we were unable to replicate this finding. This failure to replicate raises questions about the role of *GATM* in statin-related muscle injury. There is a Reply to this Brief Communication Arising by Mangravite, L. M. *et al. Nature* **513,** http://dx.doi.org/10.1038/nature13630 (2014).

Mangravite *et al.* used differential gene expression profiling of lymphoblastoid cell lines exposed to simvastatin to identify *cis*-expression quantitative trait loci (eQTLs) for the gene *GATM* as candidate loci for

Table 1 | Association of *GATM* loci with the risk of cerivastatin-related rhabdomyolysis

	All subjects					Excluding fibrate users				
SNP	Cases (n = 175), MAF	Controls (n = 645), MAF	OR	95% CI	P value	Cases (n = 76), MAF	Controls (n = 643), MAF	OR	95% CI	P value
rs9806699	0.27	0.28	1.01	0.70–1.45	0.96	0.24	0.28	0.84	0.52–1.36	0.49
rs1719247	0.29	0.25	1.37	0.98–1.90	0.07	0.24	0.25	1.00	0.64–1.57	0.99
rs1346268	0.29	0.27	1.25	0.90–1.73	0.18	0.24	0.27	0.88	0.52–1.36	0.57

Rhabdomyolysis case subjects had creatine kinase levels >10 × the upper limit of normal and used cerivastatin at the time of onset of symptoms of muscle pain or weakness. Control subjects did not experience rhabdomyolysis and used the following statins: lovastatin (44%), simvastatin (19%), pravastatin (18%), atorvastatin (13%), fluvastatin (6%) or cerivastatin (1%). CI, confidence interval; MAF, minor allele frequency; OR, odds ratio; SNP, single nucleotide polymorphism.

pharmacogenomic associations with muscle injury, which they evaluated in two case-control studies of myopathy. Variation at their most significant *cis*-eQTL for *GATM*, rs9806699, was associated with a decreased risk of muscle injury (odds ratio = 0.60, 95% confidence interval = 0.39–0.95) in a study with 72 mild myopathy cases (blood creatine kinase levels > 3 × the upper limit of normal (ULN) with muscle symptoms) recruited from a healthcare organization (Marshfield). In a second study with 39 mild and 61 severe myopathy cases (creatine kinase > 10 × ULN with muscle symptoms) using simvastatin during the SEARCH clinical trial, variation at two single-nucleotide polymorphisms (SNPs) in linkage disequilibrium with rs9806699 ($r^2 \geq 0.7$) was also associated with a decreased risk of muscle injury (rs1719247, odds ratio = 0.61, 95% confidence interval = 0.42–0.88; rs1346268, odds ratio = 0.62, 95% confidence interval = 0.43–0.90). On the basis of these epidemiologic findings and the results of functional studies in hepatocyte-derived cell lines, the authors identified *GATM* as a new genetic locus for statin-induced myopathy.

We attempted to replicate these findings in a case-control study of rhabdomyolysis (creatine kinase > 10 × ULN and muscle symptoms) related to the use of cerivastatin[3,4], which was removed from the market in 2001 because of a high incidence of this adverse drug reaction[5]. Rhabdomyolysis cases (175; 94.9% with European ancestry) were compared with statin-using control subjects from the Cardiovascular Health Study without rhabdomyolysis (645; 99.7% with European ancestry). Variation at rs9806699 was not associated with the risk of rhabdomyolysis (odds ratio = 1.01, 95% confidence interval = 0.70–1.45), and variation at the other two SNPs was weakly associated with an increased risk (rs1719247, odds ratio = 1.37, 95% confidence interval = 0.98–1.90; rs1346268, odds ratio = 1.25, 95% confidence interval = 0.90–1.73). Ninety-nine rhabdomyolysis cases used fibrates, which can cause drug–drug interactions with statins, and excluding fibrate users also resulted in null associations (Table 1). Combining our results (all subjects) with the results of Carr *et al.*[6] and Mangravite *et al.*[2] in a fixed-effects meta-analysis resulted in null associations at rs9806699 (odds ratio = 0.88, 95% confidence interval 0.72–1.08, $P = 0.22$), rs1719247 (odds ratio = 0.86, 95% confidence interval = 0.69–1.07, $P = 0.17$) and rs1346268 (odds ratio = 0.85, 95% confidence interval = 0.68–1.05, $P = 0.12$). There was statistical heterogeneity at rs1719247 ($\tau^2 = 0.22$, $P = 0.001$) and rs1346268 ($\tau^2 = 0.14$, $P = 0.009$).

Although most cases from the SEARCH trial involved severe myopathy, it is possible that the *GATM* variants identified by Mangravite *et al.* protect against mild but not severe statin-related muscle injury. Other differences in the study populations could also result in heterogeneity of the effects of these variants. An alternative explanation for the discrepant findings is that *GATM* is not related to this adverse drug reaction. By contrast, a non-synonymous variant in the drug transporter gene *SLCO1B1* (rs4149056) that decreases the clearance of statins[7,8] has been associated with statin-related muscle injury of various severity and statin types[9–12]. The odds ratio for the rs4149056 minor allele in our rhabdomyolysis study (2.0) (ref. 3) was similar to the odds ratios in a study of less-severe myopathy cases related to simvastatin use (2.1) and in a recent meta-analysis (2.2) (ref. 11). In other words, the drug transporter encoded by *SLCO1B1* is a widely replicated finding[7].

The approach by Mangravite *et al.*[2] of identifying potential new pharmacogenomic interactions through differential gene expression profiling is innovative. However, the failure to replicate their findings in a large study of rhabdomyolysis raises questions about whether *GATM* represents a genuine genetic locus for this adverse drug reaction.

Methods

Case subjects were recruited through attorneys representing cerivastatin users who developed rhabdomyolysis. Trained abstractors reviewed medical records to validate rhabdomyolysis events. As cerivastatin comprised a small fraction of statin use during its market life (March 1998 to August 2001), it was not practicable to assemble a broad sample of cerivastatin users who did not develop rhabdomyolysis. Instead, the control group comprised statin-using participants of the Cardiovascular Health Study, a prospective cohort study of older adults[13,14]. This work was funded by a grant from the NHLBI, HL078888.

James S. Floyd[1,2], Joshua C. Bis[1,2], Jennifer A. Brody[1,2], Susan R. Heckbert[1,3,4], Kenneth Rice[1,5] & Bruce M. Psaty[1,2,3,4]

[1]Cardiovascular Health Research Unit, University of Washington, 1730 Minor Avenue, Suite 1360, Seattle, Washington 98101, USA.
email: jfloyd@uw.edu

[2]Department of Medicine, University of Washington, 1959 Northeast Pacific Street, Box 356420, Seattle, Washington 98195-6420, USA.

[3]Department of Epidemiology, University of Washington, 1959 Northeast Pacific Street, Box 357236, Seattle, Washington 98195-7236, USA.

[4]Group Health Research Institute, Group Health Cooperative, 1730 Minor Avenue, Suite 1600, Seattle, Washington 98101-1448, USA.

[5]Department of Biostatistics, University of Washington, 1959 Northeast Pacific Street, Box 357232, Seattle, Washington 98195-7323, USA.

Received 3 December 2013; accepted 16 June 2014.

1. Thompson, P. D., Clarkson, P. & Karas, R. H. Statin-associated myopathy. *J. Am. Med. Assoc.* **289,** 1681–1690 (2003).
2. Mangravite, L. M., Engelhardt, B. E., Stephens, M. & Krauss, R. M. A statin-dependent QTL for GATM expression is associated with statin-induced myopathy. *Nature* **502,** 377–380 (2013).
3. Marciante, K. D. *et al.* Cerivastatin, genetic variants, and the risk of rhabdomyolysis. *Pharmacogenet. Genomics* **21,** 280–288 (2011).
4. Floyd, J. S. *et al.* A screening study of drug–drug interactions in cerivastatin users: an adverse effect of clopidogrel. *Clin. Pharmacol. Ther.* **91,** 896–904 (2012).
5. Staffa, J. A., Chang, J. & Green, L. Cerivastatin and reports of fatal rhabdomyolysis. *N. Engl. J. Med.* **346,** 539–540 (2002).
6. Carr, D. F. *et al. GATM* gene variants and statin myopathy risk. *Nature* **513,** http://dx.doi.org/10.1038/nature13628 (2014).
7. Tamraz, B. *et al.* OATP1B1-related drug–drug and drug–gene interactions as potential risk factors for cerivastatin-induced rhabdomyolysis. *Pharmacogenet. Genomics* **23,** 355–364 (2013).
8. Kameyama, Y., Yamashita, K., Kobayashi, K., Hosokawa, M. & Chiba, K. Functional characterization of SLCO1B1 (OATP-C) variants, SLCO1B1*5, SLCO1B1*15 and SLCO1B1*15+C1007G, by using transient expression systems of HeLa and HEK293 cells. *Pharmacogenet. Genomics* **15,** 513–522 (2005).
9. SEARCH Collaborative Group *et al.* SLCO1B1 variants and statin-induced myopathy–a genomewide study. *N. Engl. J. Med.* **359,** 789–799 (2008).
10. Voora, D. *et al.* The *SLCO1B1*5* genetic variant is associated with statin-induced side effects. *J. Am. Coll. Cardiol.* **54,** 1609–1616 (2009).
11. Linde, R., Peng, L., Desai, M. & Feldman, D. The role of vitamin D and SLCO1B1*5 gene polymorphism in statin-associated myalgias. *Dermatoendocrinol.* **2,** 77–84 (2010).

12. Carr, D. F. *et al. SLCO1B1* genetic variant associated with statin-induced myopathy: a proof-of-concept study using the clinical practice research datalink. *Clin. Pharmacol. Ther.* **94,** 695–701 (2013).
13. Fried, L. P. *et al.* The Cardiovascular Health Study: design and rationale. *Ann. Epidemiol.* **1,** 263–276 (1991).
14. Psaty, B. M. *et al.* Assessing the use of medications in the elderly: methods and initial experience in the Cardiovascular Health Study. The Cardiovascular Health Study Collaborative Research Group. *J. Clin. Epidemiol.* **45,** 683–692 (1992).

Author contributions J.C.B. and J.S.F. carried out the analyses. B.M.P. obtained the funding for this work. J.S.F., F.C.B., J.A.B., S.R.H., K.R. and B.M.P. contributed to the design of the analyses and the drafting and revision of the manuscript.

Competing Financial Interests B.M.P. serves on the data and safety monitoring board (DSMB) for a clinical trial of a device funded by the manufacturer (Zoll LifeCor).

doi:10.1038/nature13629

Mangravite *et al.* reply

REPLYING TO D. F. Carr *et al. Nature* **513,** http://dx.doi.org/10.1038/nature13628 (2014); J. S. Floyd *et al. Nature* **513,** http://dx.doi.org/10.1038/nature13629 (2014)

Our study[1] tested for associations of single-nucleotide polymorphisms (SNPs) at the *GATM* loci with statin-induced myopathy based on the finding that one of these SNPs (rs986699) was associated with statin-induced expression of *GATM* in a panel of human lymphoblastoid cell lines, and the fact that *GATM* encodes the enzyme responsible for synthesis of creatine, a major source of energy in skeletal muscle[1]. Significant associations with incidence of myopathy were found for rs9806699 in statin users from the Marshfield Clinic cohort. Furthermore, significant association was reported in both the Marshfield cohort and in the SEARCH clinical trial of simvastatin treatment for two SNPs in linkage disequilibrium with the index SNP (rs1719247 and rs1346268, $r^2 > 0.7$) that were genotyped in each of these groups. We have extended our meta-analysis to include the study data reported in the accompanying Comments by Carr *et al.*[2] and Floyd *et al.*[3], two studies that individually failed to replicate this association.

The original analysis was performed on data from patients who were not on fibrates in the Marshfield and SEARCH populations. This was done to mitigate the risk that a possible modest protective effect of the SNPs would be masked by the known pharmacokinetic confounding caused by concomitant use of fibrates or other drugs that promote myopathy by altering statin pharmacokinetics[4]. We have done the same for the results of Floyd *et al.* in the meta-analyses presented below, although it is notable that, based on clinical presentation and creatine kinase levels, the majority of the myopathy cases of Floyd *et al.* were considerably more severe than in the originally reported cohorts[5]. This analytical approach was not possible for the study of Carr *et al.*, since data for this subgroup were not provided. In this regard we note that because pharmacokinetic effects are major determinants of statin toxicity, the confirmation by both Carr *et al.* and Floyd *et al.* of an association of myopathy with a functional variant of the transporter gene *SLC01B1* is not representative of the power of their analyses to detect a SNP association with a modest pharmacodynamic effect.

A fixed-effects meta-analysis yielded the following *P* values: rs9806699 (Marshfield, Carr *et al.* and Floyd *et al.*), $P = 0.085$; rs1719247 (Marshfield, SEARCH and Floyd *et al.*), $P = 0.0042$; rs1346268 (Marshfield, SEARCH and Floyd *et al.*), $P = 0.0035$. Thus, the statistical significance of the initially reported association is weakened but not eliminated by the inclusion of the additional cohorts. Future efforts to replicate these findings should give consideration to heterogeneity of patient characteristics, matching of statin exposure in cases and controls, avoidance of concomitant drug use and other confounding factors, and the statistical power to detect an association of modest effect size. We agree with Carr *et al.* that the association should be assessed in more patients and hope that a larger meta-analysis will be performed. In addition, further studies will be required to determine a mechanistic basis for a contribution of *GATM* genetic variation to the risk of statin-related myopathy. This Reply is written by the subset of authors that designed and led these analyses.

Lara M. Mangravite[1], Barbara E. Engelhardt[2]†, Matthew Stephens[2,3] & Ronald M. Krauss[4]

[1]Sage Bionetworks, Seattle, Washington 98109, USA.
email: lara.mangravite@sagebase.org
[2]Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA.
[3]Department of Statistics, University of Chicago, Chicago, Illinois 60637, USA.
[4]Children's Hospital Research Institute, Oakland, California 94609, USA.
†Present address: Biostatistics and Bioinformatics Department and Department of Statistical Science, Duke University, Durham, North Carolina 27708, USA.

1. Mangravite, L. M. *et al.* A statin-dependent QTL for GATM expression is associated with statin-induced myopathy. *Nature* **502,** 377–380 (2013).
2. Carr, D. F. *et al. GATM* gene variants and statin myopathy risk. *Nature* **513,** http://dx.doi.org/10.1038/nature13628 (2014).
3. Floyd, J. S. *et al. GATM* locus does not replicate in rhabdomyolysis study. *Nature* **513,** http://dx.doi.org/10.1038/nature13629 (2014).
4. Graham, D. J. *et al.* Incidence of hospitalized rhabdomyolysis in patients treated with lipid-lowering drugs. *J. Am. Med. Assoc.* **292,** 2585–2590 (2004).
5. Marciante, K. D. *et al.* Cerivastatin, genetic variants, and the risk of rhabdomyolysis. *Pharmacogenet. Genomics* **21,** 280–288 (2011).

doi:10.1038/nature13630

EVOLUTIONARY BIOLOGY

Radiating genomes

Genome sequences and gene-expression data from representatives of five distinct lineages of African cichlid fish reveal signatures of the genomic changes that underlie the astounding cichlid diversity seen today. SEE ARTICLE P.375

CHRIS D. JIGGINS

There are more than 2,000 species of cichlid fish, the majority of which are found in three large African lakes. This species radiation is the result of somewhere between 20 million and 45 million years of evolution, although, remarkably, around 500 species found in one of these lakes, Lake Victoria, arose in only the past 100,000 years[1,2]. The amazing diversity of these fishes, and the speed with which they have evolved, makes them truly one of the natural wonders of the world. On page 375 of this issue, Brawand *et al.*[3] report five genome sequences, one from each of the major lineages of African cichlid. The data offer insight into cichlid diversification and provide a rich resource for future genomic analyses.

The species diversity of African cichlids is matched by their diversity in both ecology and morphology (Fig. 1). They occupy a huge range of ecological niches — ranging from some fairly standard fishy activities, such as eating algae or molluscs[1], through to the more bizarre, such as the scale eaters, which use their asymmetric jaws to nibble scales from the sides of other fishes[4]. Many of these forms have evolved, apparently independently, in each of the different lineages, which have undergone varying degrees of species radiation. The new reference genomes offer exciting opportunities to identify the genetic changes that led to this extraordinary diversity of morphological and ecological traits.

As well as being a resource for future studies, the sequences hold intriguing clues to the genomic changes underlying the radiation. A well-established route for evolutionary innovation is through gene duplication, which can permit existing genes to diversify and take on new functions[5]. The genomes reported by Brawand and colleagues provide evidence for a burst of gene duplications associated with species radiation. This implies that natural selection has favoured the retention of duplicate genes in African cichlids, perhaps in part owing to their role in adapting to new environments. This hypothesis is also supported by the authors' gene-expression data, which show that many of the retained duplicate genes exhibit new expression patterns. Notably, 20% of duplicate pairs have gained a completely new tissue-specific expression domain, consistent with gene duplication having led to a new gene function.

Figure 1 | Niche diversity. The differing feeding habits of cichlid fish in the three African lakes with the highest cichlid diversity — Lake Malawi, Lake Tanganyika and Lake Victoria — provide a sample of the diversity of ecological niches they occupy. Brawand *et al.*[3] present the genome sequences of one species from each of these lakes and two river-dwelling cichlids (indicated by red stars). (For photo credits, see Figure 1 of the paper[3].)

In addition to gene duplication, there is genomic evidence for accelerated evolution of protein-coding genes in the cichlids as compared with stickleback fish, which have not undergone a similarly rapid radiation and so provide a useful control group for this analysis. The accelerated evolution in cichlids was particularly striking in opsin genes, which encode proteins involved in colour vision, and in genes encoding members of the BMP signalling pathway, which influence a wide variety of developmental processes, including jaw development.

Of course, divergence in gene function can

also occur through changes in gene regulation, without a change in the protein-coding sequence, and Brawand *et al.* also addressed this. The cichlid genomes show evidence for enhanced rates of evolution in putative regulatory elements, and high evolutionary turnover in microRNAs — a class of RNA molecule that regulates gene expression. Furthermore, the genomes reveal 40 new microRNA-encoding genes that, intriguingly, show complementary patterns of expression relative to the genes they are hypothesized to regulate. This suggests that they are involved in suppressing gene expression, perhaps to stabilize and refine expression patterns that have been acquired during the radiation.

The authors attribute the great diversity of changes seen across these genomes to a period of relaxed selection that occurred early in the radiation. During this time, the selective pressures that maintained the stability of the genome were reduced, thereby allowing genetic variation to accumulate and produce subsequent diversification into the lineages we observe today. However, accelerated evolution can result either from neutral evolution due to relaxed selection, or from positive natural selection acting through new selective pressures. Most of the genomic signatures in the paper do not strongly distinguish between these two possibilities. Indeed, it seems most likely that the retention of gene duplicates and rapid genetic divergence were primarily driven by positive natural selection, as species adapted to the great diversity of ecological niches available in the lakes. Subsequent extinction of early lineages could have led to an apparent burst of rapid change on the branch leading to the extant species. There may be no need to invoke a genetic revolution when plain old natural selection can explain the observed patterns.

Although the five genomes offer some impressive insights into cichlid biology, I believe that the most exciting advances will come from analysis of more-closely related genomes within each radiation. The cichlids offer in abundance two of the characteristics that have facilitated analysis of adaptive traits in other taxa: there are many closely related species that show highly divergent morphology, and there is repeated evolution of similar traits in parallel. Whole-genome sequencing of multiple individual fishes with both divergent and convergent ecological traits will provide rich pickings for understanding how genetic changes are associated with specific ecological characteristics. Brawand *et al.* have scratched the surface of this task by reanalysing sequence data from samples of six species found in Lake Victoria[6]; these suggest that even very closely related species show quite high levels of divergence across the genome.

These genomes will facilitate further studies that will undoubtedly enhance our understanding of cichlid biology. It may be rash, but I will make one prediction. Work in organisms ranging from sticklebacks[7] to butterflies[8] has shown that recent adaptive events can make use of ancient genetic variants. This may be surprising, but can occur because gene flow within a species, or sometimes even between species[8], can provide 'pre-adapted' variants that permit populations to adapt rapidly to new challenges. So I predict that similarities between cichlids in different lakes that are currently considered to have evolved independently will in fact turn out to have resulted in part from ancient shared variation that may have arisen early in the radiation[9]. ■

Chris D. Jiggins *is in the Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, UK.*
e-mail: c.jiggins@zoo.cam.ac.uk

1. Wagner, C. E., Harmon, L. J. & Seehausen, O. *Nature* **487,** 366–369 (2012).
2. Genner, M. J. *et al. Mol. Biol. Evol.* **24,** 1269–1282 (2007).
3. Brawand, D. *et al. Nature* **513,** 375–381 (2014).
4. Lee, H. J., Kusche, H. & Meyer, A. *PLoS ONE* **7,** e44670 (2012).
5. Ohno, S. *Evolution by Gene Duplication* (Springer, 2014).
6. Wagner, C. E. *et al. Mol. Ecol.* **22,** 787–798 (2013).
7. Colosimo, P. F. *et al. Science* **307,** 1928–1933 (2005).
8. The Heliconius Genome Consortium. *Nature* **487,** 94–98 (2012).
9. Loh, Y.-H. E. *et al. Mol. Biol. Evol.* **30,** 906–917 (2013).

This article was published online on 3 September 2014.

CONDENSED-MATTER PHYSICS

Catching relativistic electrons

Low-energy electrons have been found to mimic relativistic high-energy particles in cadmium arsenide. This defines the first stable '3D Dirac semimetal', which holds promise for fundamental-physics exploration and practical applications.

ZHIHUAI ZHU & JENNIFER E. HOFFMAN

In classical Newtonian mechanics, an object's energy varies as the square of its velocity or momentum (Fig. 1a) — a rule that car drivers should treat with respect. Photons, neutrinos and other light, fast-moving particles are governed instead by Einstein's theory of relativity: their energy scales linearly with their momentum, with fixed velocity equal to the slope of the increase. Such relativistic high-energy particles hold the key to fundamental understanding of our Universe. But where do electrons — which determine the more practical properties of the materials immediately around us — fit into this picture? Electrons move very fast, but their motion is not primarily relativistic in conventional solids. However, in a paper published in *Physical Review Letters*, Borisenko *et al.*[1] report the discovery of relativistic motion of low-energy electrons in cadmium arsenide (Cd_3As_2). Taken together with similar findings described in three independent papers, by Neupane *et al.*[2], Liu *et al.*[3] and Jeon *et al.*[4], this result paves the way for future relativistic electronics.

The realization that low-energy electrons can mimic high-energy relativistic particles occurred a decade ago with the isolation of two-dimensional (2D) carbon in the form of graphene[5]. This material has dual significance for the exploration of fundamental physics and for revolutionary applications; it has prompted more than 100,000 publications, some 7,000 patent applications and a 2010 Nobel prize. Electrons in graphene are described as massless Dirac fermions because they have half-integer spin, which makes them fermions, and their linear energy–momentum relationship obeys Dirac's famous wave equation, which first united quantum mechanics and special relativity almost a century ago. Graphene is also a semimetal, meaning that its Fermi energy (the dividing line between filled and empty electronic states) sits ideally at its 'Dirac point' — where its valence and conduction energy bands meet (Fig. 1b) — and may be easily tuned using an applied voltage. The resultant charge carriers may be either electrons or holes (the absence of electrons) and have high mobility: a measure of inverse electrical resistivity per carrier, which increases with carrier velocity but decreases with carrier scattering.

Graphene's moderately high carrier velocity of about 10^5 metres per second, combined with the reduced intrinsic scattering possibilities caused by the small carrier density inherent to a Dirac semimetal, can give a mobility up to 140 times that of silicon — the material of choice for most electronic applications. Therefore, graphene offers promise for making novel, high-efficiency electronic devices. However, graphene is challenging to fabricate and manipulate in large sheets, and its mobility is extremely susceptible to scattering from environmental

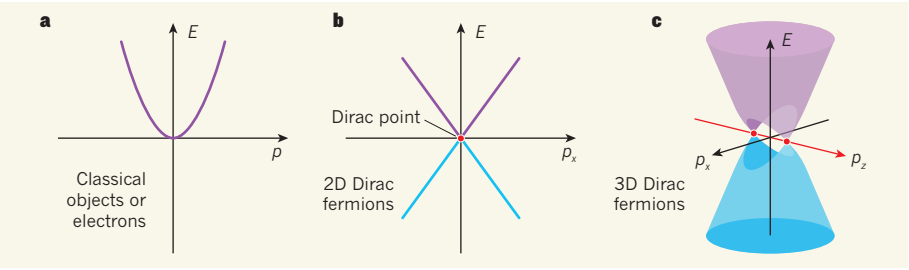


Figure 1 | Energy–momentum spectra of electrons. **a**, Classical objects and electrons exhibit a parabolic relationship between their energy (E) and momentum (p). **b**, Two-dimensional (2D) Dirac fermions, such as electrons in graphene, have valence (blue) and conduction (purple) energy bands with a linear energy–momentum relationship. These touch at a point called the Dirac point in the 3D parameter space formed by E, p_x and p_y. Shown here is a 2D slice of the 3D space. **c**, A 3D slice of the 4D (E, p_x, p_y, p_z) energy–momentum relationship of 3D Dirac fermions such as those discovered[1–4] in Cd_3As_2, with two Dirac points along a special high-symmetry axis (p_z).

defects because graphene is all surface.

A second kind of 2D Dirac semimetal arises from another relativistic effect of electrons called spin–orbit coupling — the interaction between an electron's spin and the induced magnetic field from the electron's orbital motion. Spin–orbit coupling is generally small for materials that are made up of light atoms such as carbon, but for materials containing heavy atoms such as bismuth and cadmium, the interaction can be significant; for example, it can invert the valence and conduction bands in the bulk of an insulator. This inversion can lead to surface Dirac fermions that are topologically protected — surface carriers that are robust against some local disorder and have their spin locked to their momentum (that is, the carrier's momentum determines its spin).

These 'topological insulators'[6,7] provoked tremendous excitement in recent years about possible applications such as low-energy-consumption spintronic devices, which manipulate the spin rather than the charge of electrons, for high-performance computing. But despite their name, existing topological insulators have excess conducting bulk electrons, which overwhelm the surface Dirac fermions and foil their use.

Meanwhile, new ideas were brewing, suggesting that 3D Dirac semimetallic states could exist in the bulk of a solid material. It was known that such states could occur under finely tuned conditions, such as the exact concentration of bismuth at which spin–orbit coupling becomes strong enough to invert the bulk energy bands in antimony–bismuth alloys[8] ($Sb_{1-x}Bi_x$). But more recent theoretical work predicted the robust occurrence of such states in pure materials that have certain crystalline symmetries: first, unstable BiO_2 (ref. 9), then air-sensitive Na_3Bi (ref. 10) and, finally, the stable compound Cd_3As_2 (ref. 11). Furthermore, when time-reversal or spatial-inversion symmetries are broken — for example, by application of a magnetic field or pressure — each Dirac point can split into two copies at which the electrons become Weyl fermions[12], which have opposite chirality (spin orientation with respect to their direction of motion). These Weyl fermions could enable robust spintronics in three dimensions.

Cd_3As_2 has been known for more than 50 years[13] for its extraordinary carrier mobility, which is larger than that of suspended graphene and among the highest of any bulk semiconductor. Thanks to the recent studies by Borisenko *et al.*[1], Neupane *et al.*[2] and Liu *et al.*[3] — who all conducted experiments on Cd_3As_2 using a technique called angle-resolved photoemission spectroscopy (ARPES) — we now understand that the high mobility arises from high-velocity 3D Dirac semimetallic states.

During ARPES experiments, monochromatic light is incident on a sample and electrons can absorb a photon and escape from the material. To unveil the full 3D energy–momentum relationship of electrons within Cd_3As_2, a challenging but crucial step was to precisely measure the energy and momentum of emitted electrons while tuning the photon energy through a wide range. The data[1–3] clearly show a linear energy–momentum relationship, with two Dirac points along a crystal axis of four-fold rotational symmetry (Fig. 1c). This result proves that electrons in this material are 3D massless Dirac fermions as predicted[11]. Measurement of the energy–momentum slope gives electron velocity as high as about 10^6 m s^{-1} (ref. 2), but with a ten-fold discrepancy between the three studies[1–3], which could be due to differences in sample quality or the angle of the exposed surface. Liu *et al.* additionally demonstrated that the carrier concentration in Cd_3As_2 could be finely tuned by 'doping' the surface of the material with potassium atoms[3], making it a flexible platform for future studies.

Most recently, Jeon *et al.*[4] used a scanning tunnelling microscope to confirm Cd_3As_2 as a 3D Dirac semimetal down to atomic length scales, and to visualize how dopant atoms scatter carriers primarily in the valence band, preserving the mobility of carriers in the high-velocity conduction band. Furthermore, Jeon and colleagues applied a magnetic field, which is not possible in an ARPES experiment. Although the field breaks the time-reversal symmetry that would be necessary to split the Dirac fermions into the more exotic chiral Weyl fermions, its orientation in this experiment also breaks the four-fold rotational symmetry of the crystal that was necessary to realize the Dirac fermions in the first place. This means that the first glimpse of Weyl fermions will need to wait for a follow-up experiment in which the magnetic field has a different orientation.

The work on Cd_3As_2 (refs 1–4), together with the lower-mobility Na_3Bi reported earlier this year[14], confirms the existence of motion of Dirac fermions inside 3D materials. Despite its exciting new physics, the application potential of Cd_3As_2 is limited by its small band-inversion energy — the relativistic nature is not robust at room temperature[4]. Furthermore, Cd_3As_2 is not exactly something you want in your drinking water. Nevertheless, given the new understanding that robust Dirac fermions can arise in solids from general crystalline symmetries and strong spin–orbit coupling, there are probably numerous 3D Dirac semimetals yet to be discovered[9]. Immediate research priorities include magnetic-field and pressure control to isolate chiral Weyl fermions in existing materials, realization of these materials as thin films to access a phenomenon known as the quantum spin Hall effect to visualize the spatial flow of surface Dirac fermions[15], and computational modelling to predict new materials and heterostructures with larger band-inversion energies[16]. Then exotic applications, such as a 'chiral battery' or a 'quantum amplifier' of magnetic field, may be on the horizon[17]. ■

Zhihuai Zhu *and* **Jennifer E. Hoffman**
are in the Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA.
e-mails: zhzhu@physics.harvard.edu; jhoffman@physics.harvard.edu

1. Borisenko, S. *et al. Phys. Rev. Lett.* **113,** 027603 (2014).
2. Neupane, M. *et al. Nature Commun.* **5,** 3786; http://dx.doi.org/10.1038/ncomms4786 (2014).
3. Liu, Z. K. *et al. Nature Mater.* **13,** 677–681 (2014).
4. Jeon, S. *et al. Nature Mater.* **13,** 851–856 (2014).
5. Geim, A. K. & Novoselov, K. S. *Nature Mater.* **6,** 183–191 (2007).
6. Hasan, M. Z. & Kane, C. L. *Rev. Mod. Phys.* **82,** 3045 (2010).
7. Qi, X.-L. & Zhang, S.-C. *Rev. Mod. Phys.* **83,** 1057 (2011).
8. Hsieh, D. *et al. Nature* **452,** 970–974 (2008).
9. Young, S. M. *et al. Phys. Rev. Lett.* **108,** 140405 (2012).
10. Wang, Z. *et al. Phys. Rev. B* **85,** 195320 (2012).
11. Wang, Z., Weng, H., Wu, Q., Dai, X. & Fang, Z. *Phys. Rev. B* **88,** 125427 (2013).
12. Wan, X., Turner, A. M., Vishwanath, A. & Savrasov, S. Y. *Phys. Rev. B* **83,** 205101 (2011).
13. Rosenberg, A. J. & Harman, T. C. *J. Appl. Phys.* **30,** 1621 (1959).
14. Liu, Z. K. *et al. Science* **343,** 864–867 (2014).
15. Kane, C. L. & Mele, E. J. *Phys. Rev. Lett.* **95,** 226801 (2005).
16. Burkov, A. A. & Balents, L. *Phys. Rev. Lett.* **107,** 127205 (2011).
17. Kharzeev, D. E. & Yee, H.-U. *Phys. Rev. B* **88,** 115119 (2013).

ANIMAL BEHAVIOUR

The evolutionary roots of lethal conflict

A comprehensive analysis of lethal coalitionary aggression in chimpanzees convincingly demonstrates that such aggression is an adaptive behaviour, not one that has emerged in response to human impacts. SEE LETTER P.414

JOAN B. SILK

In 2013, there were 33 armed state-level conflicts around the world[1]. Many of these had persisted for decades, killed thousands of people and thwarted international peacekeeping efforts. War is certainly a contemporary fixture, but has it always been one? There is vigorous disagreement over the answer to this question. Some argue that warfare has been a pervasive feature throughout human history and has had important effects on human nature[2], whereas others contend that war is rare in foraging groups[3], the kinds of societies that we lived in for most of our evolutionary history. Debates about the origins and prevalence of human warfare are echoed in the question of whether lethal coalitionary aggression in chimpanzees has evolved through natural selection or whether it is a non-adaptive consequence of human disturbance. In this issue, Wilson *et al.*[4] (page 414) argue persuasively on the side of adaptation.

Many species of non-human primates have hostile relationships with members of neighbouring groups, and some species collectively defend the boundaries of their territories. But intergroup encounters rarely lead to serious injuries or deaths, perhaps because the risks of escalated aggression usually do not outweigh the benefits of killing opponents. Lethal coalitionary attacks on individuals from neighbouring communities have been documented only in chimpanzees. The first report of such killings was published 35 years ago[5], but the debate about their adaptive significance continues.

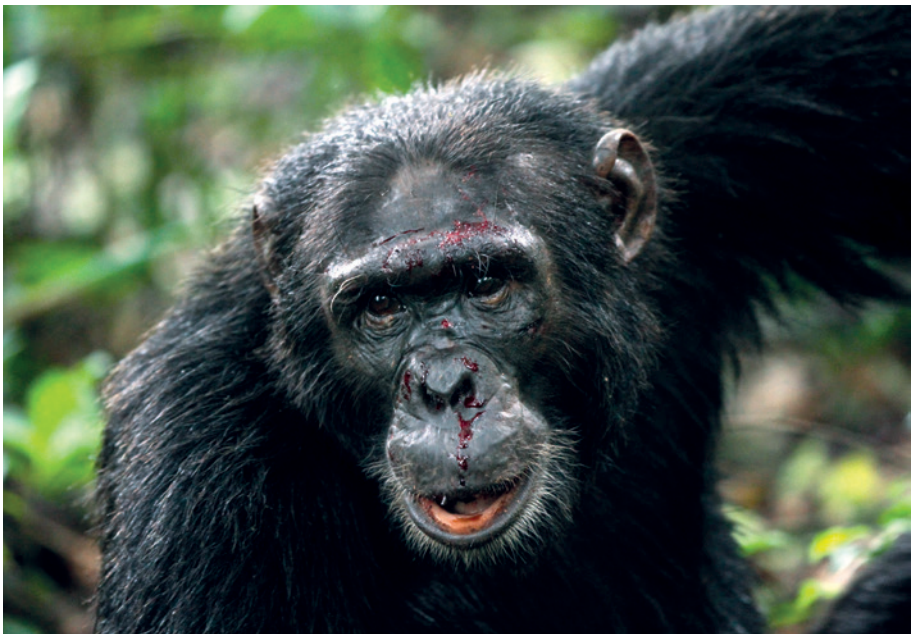
One point of view is that natural selection has favoured the evolution of lethal coalitionary intergroup aggression in chimpanzees as a means to enhance access to valuable resources, such as food and mates. Intergroup aggression might be more deadly in chimpanzees than in most other species because chimpanzees can exploit the imbalances of power that arise from 'fission–fusion' social organization[6]. Chimpanzees often fragment into temporary parties that travel and forage independently within their community's home range. When parties of males encounter single individuals from other communities, they sometimes launch brutal assaults that leave victims gravely wounded or dead (Fig. 1).

The opposing view is that lethal aggression is a non-adaptive response to anthropogenic influences, particularly artificial provisioning. In early primate field studies, researchers often used food to lure animals out of the forest, to facilitate close-range observation and to enhance habituation. At the Gombe National Park in Tanzania, where in 1962 primatologist Jane Goodall began providing bananas to chimpanzees who visited her camp, this practice had the desired effect. However, the chimpanzees began spending more and more time in the camp and rates of aggression among them increased. Provisioning was curtailed, and eventually terminated altogether. There were no reports of lethal aggression at Gombe before provisioning began, leading some researchers to conclude that these killings were the consequence of human intervention[7]. Subsequent reports of killings at other sites, where chimpanzees had never been provisioned, have been attributed to other forms of human intervention, including habitat loss[8].

Wilson *et al.* have comprehensively tested these two hypotheses by analysing 426 combined years of research at 18 chimpanzee (*Pan troglodytes*) study sites, and 92 years of research at 4 bonobo (*Pan paniscus*) study sites. The authors assembled information on all instances of lethal aggression that have been observed directly by researchers, inferred from the nature of the victim's injuries or suspected on the basis of the circumstances of their deaths or disappearances. Coalitional killings were documented at 15 of the 18 chimpanzee study sites, but there was only one suspected killing among bonobos.

The authors then tested how the frequency of killings by chimpanzees was affected by several variables linked to human impact, including provisioning and habitat disturbance, and a second set of variables related to the intensity of resource competition, including the number of males and population density. The statistical model that best fits the data includes variables linked to the intensity of resource competition, rather than those linked to human impact. Specifically, the authors' modelling shows that killings occur at higher rates in communities that have more males and higher population densities.

These results should finally put an end to the idea that lethal aggression in chimpanzees is a non-adaptive by-product of anthropogenic influences — but they will probably not be enough to convince everyone. Perceptions of the behaviour of non-human primates, particularly chimpanzees, are often distorted by ideology and anthropomorphism, which produce a predisposition to believe that morally desirable features, such as empathy and altruism, have deep evolutionary roots, whereas undesirable features, such as group-level violence and sexual coercion, do not. This reflects a naive form of biological determinism. Selective pressures alter traits as organisms move into new environments and confront new



ANDREW BERNARD

Figure 1 | A male chimpanzee with fresh wounds following an inter-group attack.

challenges and opportunities. The data tell us that there are some ecological and demographic circumstances in which the benefits of lethal aggression exceed the costs for chimpanzees, nothing more. Humans are not destined to be warlike because chimpanzees sometimes kill their neighbours.

For those who are persuaded by Wilson and colleagues' evidence, a more interesting set of questions emerges. For example, how do chimpanzees overcome the collective-action problem? By eliminating rival males and infants sired by males from other communities, chimpanzees gain access to new territories and mating partners. But these benefits flow to the group as a whole, which creates opportunities for free-riding. Although the imbalance-of-power hypothesis relies on the odds being in the aggressors' favour, males forgo opportunities to mate and forage while they are on patrol, and run at least some risk of being injured in attacks. Do males that join patrols and lead attacks gain more benefits than those that remain in the security of their own community's territory? What forces curtail free-riding? The answers to these questions will provide interesting insight into the selective forces that favour group-level cooperation in species without language, social institutions and systems for sanctioning free-riders. ■

Joan B. Silk *is in the School of Human Evolution & Social Change, Arizona State University, Tempe, Arizona 85287-2402, USA.*
e-mail: joan.silk@asu.edu

1. Themnér, L. & Wallensteen, P. *J. Peace Res.* **51,** 541–554 (2014).
2. Choi, J.-K. & Bowles, S. *Science* **318,** 636–640 (2007).
3. Fry, D. P. & Söderberg, P. *Science* **341,** 270–273 (2013).
4. Wilson, M. L. *et al. Nature* **513,** 414–417 (2014).
5. Goodall, J. *et al.* in *The Great Apes* (eds Hamburg, D. A. & McCown, E. R.) 13–53 (Benjamin/Cummings, 1979).
6. Crofoot, M. C. & Wrangham, R. W. in *Mind the Gap: Tracing the Origins of Human Universals* (eds Kappeler, P. M. & Silk, J. B.) 171–195 (Springer, 2010).
7. Power, M. *The Egalitarians — Human and Chimpanzee: An Anthropological View of Social Organization* (Cambridge Univ. Press, 2005).
8. Ferguson, R. B. in *Origins of Altruism and Cooperation* (eds Sussman, R. W. & Cloninger, C. R.) 249–270 (Springer, 2011).

ASTROPHYSICS

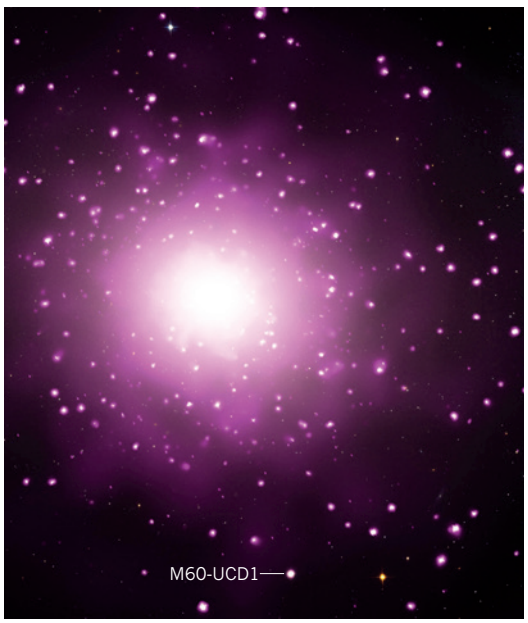
Giant black hole in a stripped galaxy

An oversized, supermassive black hole has been discovered at the centre of a densely packed conglomeration of stars. The finding suggests that the system is the stripped nucleus of a once-larger galaxy. SEE LETTER P.398

AMY E. REINES

Supermassive black holes, which have masses millions or even billions of times that of our Sun, reside at the centre of almost every massive galaxy, including our Milky Way[1]. There seems to be some connection between the evolution of galaxies and that of these black holes, although the nature of the relationship is not well understood. What we do know is that, in general, bigger galaxies harbour bigger black holes at their centres, and these black holes are typically about 0.5% of the total mass of a spheroidal galaxy's stars[1]. But on page 398 of this issue, Seth *et al.*[2] report the detection of an oversized supermassive black hole that is a whopping 18% of the stellar mass of its unusual host.

The dense stellar system in which the black hole has been found, M60-UCD1 (ref. 3), is called an ultra-compact dwarf galaxy, and marks a previously unknown environment for supermassive black holes. Ultra-compact dwarf galaxies are densely packed spherical conglomerations of stars[4]. For years, astronomers have debated the nature of these objects — are they extremely massive star clusters, or are they the nuclei of galaxies that have had their outer layers stripped off through gravitational interactions with other galaxies? Seth and colleagues present the first clear case that an individual ultra-compact dwarf is a stripped-galaxy nucleus, because star clusters do not host supermassive black holes.



X-RAY: NASA/CXC/MSU/J. STRADER *ET AL.*; OPTICAL: NASA/STSCI

Figure 1 | Dwarfed by its neighbour. This composite image, constructed from data from NASA's Chandra X-ray Observatory and Hubble Space Telescope, depicts the massive elliptical galaxy M60 and the nearby ultra-compact dwarf galaxy M60-UCD1. Seth *et al.*[2] report that M60-UCD1 contains a supermassive black hole.

Seth *et al.* 'weighed' the black hole by determining its gravitational influence on nearby stars orbiting it[5,6]. To explain the observed stellar velocities and the distribution of light within M60-UCD1, they had to invoke the presence of a central black hole with a mass 21 million times that of our Sun. A black hole of that size would be expected to reside in a host galaxy with a mass of about 7 billion solar masses. However, Seth *et al.* estimate that M60-UCD1 has a stellar mass of only 120 million solar masses.

Although the discovery of an enormous black hole in such a small galaxy is surprising, recent work has uncovered a substantial number of black holes in other low-mass dwarf galaxies[7]. However, M60-UCD1 is clearly a different beast from those — it is far more compact and has a much more massive black hole. The small black holes in other low-mass dwarf galaxies are probably similar to the first 'seeds' of supermassive black holes[8]. Over cosmic time, such seeds grow by swallowing gas and coalescing with other black holes during galaxy mergers. With a mass 200 times that of the smallest nuclear black holes known, M60-UCD1's black hole seems to have already grown considerably.

So how did such a big black hole get into such a tiny galaxy? The answer may be related to M60-UCD1's galactic neighbourhood. This ultra-compact dwarf galaxy is right next door to the giant elliptical galaxy M60 (Fig. 1). Seth and co-workers' simulations show that M60-UCD1 may have formerly been a more massive galaxy than it is now (with a proportionally sized black hole), but lost most of its stars in a gravitational tug of war while orbiting its giant neighbour. What is left today is the dense stellar nucleus and central supermassive black hole from the larger progenitor galaxy.

The evidence for a supermassive black hole in M60-UCD1 is strong, but it is not the only possible explanation for the

data. Although seemingly unlikely, we cannot definitively rule out the existence of a population of low-mass stars or dead stellar remnants at the galaxy's centre that do not produce much visible light. An X-ray source that might be produced by a supermassive black hole has been detected at the heart of M60-UCD1, but this radiation could also be generated by the dead remnant of a star[3]. Follow-up observations at radio wavelengths could distinguish between these possibilities[9,10] and provide further support for the presence of a supermassive black hole.

Seth and colleagues' discovery is an important step towards understanding the nature of ultra-compact dwarf galaxies. Many other ultra-compact dwarfs show tantalizing hints that they, too, harbour supermassive black holes and are therefore stripped galaxy nuclei, but direct evidence is lacking. The authors are participating in ongoing observing programmes that may provide conclusive evidence for supermassive black holes in four other ultra-compact dwarfs. But at present, detecting the gravitational pull of a black hole on surrounding stars is feasible for only the brightest and closest systems. Attempting to detect the direct gravitational signatures of black holes in a large population of ultra-compact dwarfs must therefore wait for the next generation of telescopes.

If supermassive black holes are indeed commonplace in ultra-compact dwarfs, this would have major implications for the demographics of such black holes — Seth *et al.* estimate that there could be more than double the number of supermassive black holes in the local Universe than is presently thought. Although this is possible, it is far from certain. Future studies will tell us whether M60-UCD1 is a fluke, or whether other ultra-compact dwarfs are also stripped galactic nuclei that host black holes. ■

Amy E. Reines *is in the Department of Astronomy, University of Michigan, Ann Arbor, Michigan 48109-1107, USA.*
e-mail: reines@umich.edu

1. Kormendy, J. & Ho, L. C. *Annu. Rev. Astron. Astrophys.* **51,** 511–653 (2013).
2. Seth, A. C. *et al. Nature* **513,** 398–400 (2014).
3. Strader, J. *et al. Astrophys. J.* **775,** L6 (2013).
4. Brodie, J. P., Romanowsky, A. J., Strader, J. & Forbes, D. A. *Astron. J.* **142,** 199 (2011).
5. Schwarzschild, M. *Astrophys. J.* **232,** 236–247 (1979).
6. van den Bosch, R. C. E. & de Zeeuw, P. T. *Mon. Not. R. Astron. Soc.* **401,** 1770–1780 (2010).
7. Reines, A. E., Greene, J. E. & Geha, M. *Astrophys. J.* **775,** 116 (2013).
8. Volonteri, M. *Astron. Astrophys. Rev.* **18,** 279–315 (2010).
9. Gültekin, K., Cackett, E. M., King, A. L., Miller, J. M. & Pinkney, J. *Astrophys. J.* **788,** L22 (2014).
10. Merloni, A., Heinz, S. & Di Matteo, T. *Mon. Not. R. Astron. Soc.* **345,** 1057–1076 (2003).

NEUROSCIENCE

Shedding light on a change of mind

Sophisticated genetic tools that make brain cells responsive to light have now been used in mice to trigger a memory connected with a particular place, and to switch its association from negative to positive, or vice versa. SEE LETTER P.426

TOMONORI TAKEUCHI & RICHARD G. M. MORRIS

We often believe that our memories are accurate, but in fact they can be malleable, changing over time as recollections become less precise or as events that never happened are falsely remembered[1]. There is also another way in which memory can change. The memory of a romantic first meal out with a partner may take on a different mood when the relationship falters. That of a favourite family beach in summer may be destroyed by witnessing a swimming tragedy there. In these cases, memory of the place remains accurate, but the positive associations with that place are lost. On page 426 of this issue, Redondo *et al.*[2] investigate the neural basis of this selective change.

Our memories are representations of past experiences that are believed to be encoded in networks of neurons that fire together or in sequence. The representation of a particular place — a 'where' memory — is encoded in a brain structure called the hippocampal formation, which is embedded within the medial temporal lobe. A separate representation in the amygdala of the brain encodes a 'what' memory, which recalls whether one feels good about a place (a positive valence) or has marked it off as dangerous (a negative valence). These two representations are thought to become connected during learning. The amygdala also has direct downstream connections to the action and endocrine systems that are involved in approach and avoidance[3].

Redondo and colleagues investigated the separate representations of 'where' and

50 Years Ago

Journey to the Jade Sea. By John Hillaby — Books by writers who go to Africa in search of their souls are always interesting to us who went there in search of wages; this book is fascinating. "Essentially, I walked into the N.F.D. for the hell of it" (p. 2); Mr. Hillaby took his own hell by using a small string of unhealthy camels for transport instead of a lorry or Land-Rover, as do other people in that part of Kenya. "Perhaps all safaris start this way. Somewhat despondently …" (p. 7); they do not, but I should have been despondent if I had started with that collection of provisions … in old cardboard boxes … Messrs. Constable have published a most interesting book. They might also publish an interesting one by the Warden, for he no doubt would tell us more about the animals and plants.
From *Nature* 19 September 1964

100 Years Ago

Let us consider lastly a disease which collects the last toll from one-seventh of humanity, and debilitates and enfeebles the lives of many whom it does not entirely destroy … How are we organizing our campaign against tuberculosis? Bacteriology has taught us that it is an infectious disease and has isolated the organism … all over the civilized world the total death-roll of human kind annually from tuberculosis probably does not fall short of a million souls … This disease must be stopped at its source as well as dealt with on its course. No disease has ever been eradicated from a community by discovering cures for it, and none ever will; many diseases have disappeared because their sources have been cut off. Let us be scientific, let us search out the truth; having found it, let us act upon it, and let us conceal nothing that is true.
From *Nature* 17 September 1914

'what' memories, and looked at whether the associations between them could be changed. To do this, they used several molecular-engineering tools, including optogenetics, which enables the manipulation of neurons in response to light. The authors genetically engineered male mice such that, when the antibiotic doxycycline was removed from their daily diet, a light-sensitive protein called channelrhodopsin-2 (ChR2) was able to be expressed. Depending on the genetic tools used, this took place in neurons of either the hippocampus or the amygdala in which the gene *c-fos* was active — a response to neural activity and learning. When doxycycline was briefly removed from the diet and the animals were given either contextual fear conditioning (a small electric shock) or reward conditioning (interaction with a female mouse), memory encoding caused *c-fos* activation in these brain areas and resulted in labelling of the 'where' or 'what' memory neurons with ChR2.

After training, the authors added doxycycline to the diet of the mice once again, preventing any further ChR2 labelling and ensuring that only the training memory representations could be activated by blue light. Redondo and co-workers then performed a place-preference test, in which blue light was turned on whenever the mice entered a designated target zone. This selectively activated the ChR2-labelled neurons and the networks to which they had been associatively connected. Fear-conditioned mice duly moved away from the target zone, because the negative memory was optogenetically reactivated when they were in this area, whereas reward-conditioned mice stayed longer in the target zone, recalling the positive memory.

The key aim of the study was to determine whether it is possible to change the 'what' association linked to a 'where' representation. The authors were able to do this, but the additional conditioning did not involve returning the animals to the training arena. Instead, the researchers optogenetically reactivated the appropriate hippocampal neurons of fear-conditioned mice while allowing them access to a female. The outcome was a successful switch of the original aversive association with the target zone into an attractive association. A switch from attraction to fear could also be achieved. However, the amygdala representations of aversion or attraction stayed as they were — their downstream connections were unchanged.

These data imply that functional connectivity between memory representations in the hippocampus and the amygdala can be altered. Analysing these connections under the microscope, Redondo *et al.* found that additional conditioning led to decreases in the proportion of hippocampal-activated neurons in the amygdala that represented the original conditioning. Finally, they observed that if a hippocampal representation was formed by fear conditioning, and then had its valence changed through subsequent conditioning with reward, the animals would later display less fear when returned to the original fear-conditioning box.

It has long been apparent that memories can be changed from bad to good, or vice versa. What is so intriguing about this study is that the memory representations associated with a place are dissected into their network components and, rather than re-exposing the animals to the training situation to achieve a change, light is used to selectively reactivate the representation of the 'where' component of a memory and then change its 'what' association.

Contemporary theories of learning are less about stimuli and responses than about the internal representation of events. For example, work done in the past 10 years has focused on associatively activated event representations in learning; that is, on the acquisition of memories that can later be evoked by a reminder cue of a specific stimulus or by the act of returning to a particular place[4]. A key finding of this work was that associatively activated event representations can successfully substitute for the events that created them, with the possibility that new learning will successfully associate new information to the event memory evoked by the reminder cue. This may result in an altered response to that cue.

Optogenetic techniques, used so ingeniously by Redondo and colleagues, complement and expand on this previous work[4] and on the classical 'disconnection' approach, which involves unilaterally damaging two structures on opposite sides of the brain to establish the importance of their anatomical connectivity for learning[5,6]. The use of ChR2 cell labelling in a way that is temporally controlled and dependent on neural activity, followed by optogenetic reactivation of the representation, takes us closer to identifying the networks that underlie certain forms of memory.

There are limitations to this optogenetic strategy, notably when sequences of neural firing are the essence of the memory[7] (for instance, the memory of a musical tune or a sequence of actions). This is because the resulting labelling will represent the sum of all the neurons that upregulate the activity-regulated genes, such as *c-fos*, rather than any explicit representation of sequence. Through optogenetics and exceptionally careful design of behavioural studies, molecular engineering is nonetheless shedding light on our understanding of the underlying physiological networks of memory. ■

Tomonori Takeuchi *and* **Richard G. M. Morris** *are at the Centre for Cognitive and Neural Systems, University of Edinburgh, Edinburgh EH8 9JZ, UK.*
e-mails: tomonori.takeuchi@ed.ac.uk; r.g.m.morris@ed.ac.uk

1. Loftus, E. F. & Palmer, J. C. *J. Verbal Learn. Verbal Behav.* **13,** 585–589 (1974).
2. Redondo, R. L. *et al. Nature* **513,** 426–430 (2014).
3. LeDoux, J. E. *Proc. Natl Acad. Sci. USA* **111,** 2871–2878 (2014).
4. Saddoris, M. P., Holland, P. C. & Gallagher, M. *J. Neurosci.* **29,** 15386–15396 (2009).
5. Ettlinger, G. *Brain* **82,** 232–250 (1959).
6. Gaffan, D. & Wilson, C. R. E. *Cortex* **44,** 928–935 (2008).
7. Xu, S., Jiang, W., Poo, M.-M. & Dan, Y. *Nature Neurosci.* **15,** 449–455 (2012).

This article was published online on 27 August 2014.

ORGANIC CHEMISTRY

Reactivity tamed one bond at a time

A catalyst that couples together three reactants to form just one compound out of several possibilities, as a single mirror-image isomer, should simplify the synthesis of biologically relevant molecules. **SEE ARTICLE P.367**

MATTHEW T. VILLAUME & PHIL S. BARAN

When the structural complexity of a molecule reaches a certain point, accessing it in a cheap, high-yielding and short chemical synthesis can seem impossible. As a result, interesting target molecules remain on the blackboard, never to be properly studied. Organic chemists are trying to deal with this by developing methods that quickly assemble complex molecules. A process called multicomponent coupling is at the forefront of this work, but it requires the reactivity of many different molecules to be tamed simultaneously, rather like orchestrating a three-ring circus. On page 367 of this issue, Meng *et al.*[1] report remarkable progress in this compelling area of research.

The chemical synthesis of structurally complex molecules, such as high-value compounds for the pharmaceutical, agricultural and materials industries, can be laborious

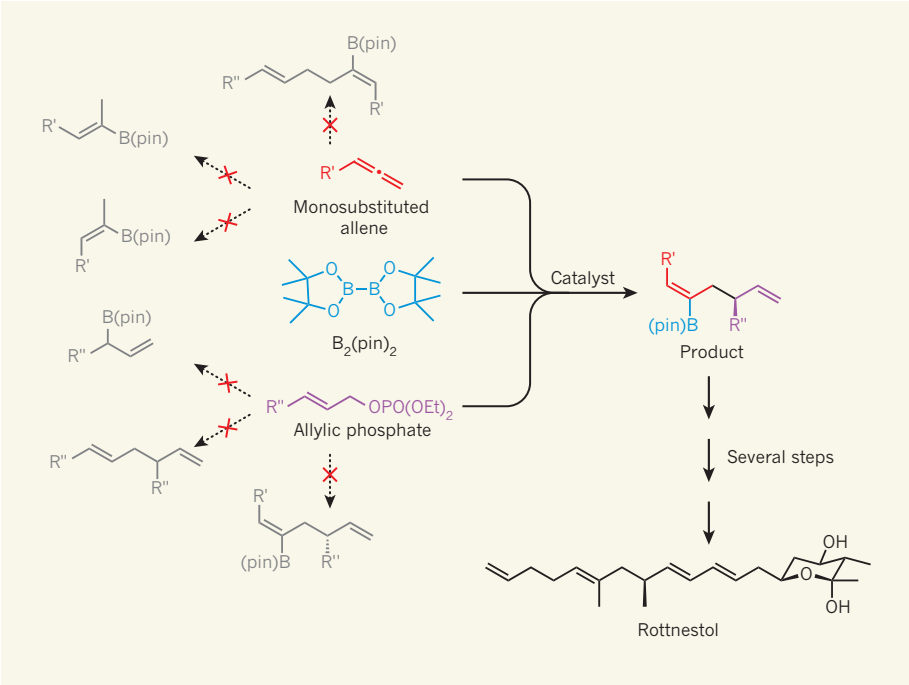


Figure 1 | One catalyst rules all. Meng *et al.*[1] report a catalyst that controls the reaction of three compounds — a monosubstituted allene, bis(pinacolato)diboron ($B_2(pin)_2$) and an allylic phosphate — to form a single product out of many possible compounds. They used this product as a key intermediate in the synthesis of the naturally occurring antibiotic rottnestol. R′ and R″ represent different attached chemical groups; Et, ethyl; pin, pinacolato ($OC(CH_3)_2C(CH_3)_2O$).

and time consuming for even the most skilled chemist[2]. Each bond and group in a molecule adds exponentially to the time and care needed to produce it. Multicomponent coupling reactions help to solve this problem by joining together more than two reagents sequentially in a single reaction vessel[3].

These kinds of reactions have been studied for more than 100 years, with early work focused on the preparation of heterocyclic compounds (rings containing more than one kind of atom) through simple 'condensation' chemistry[4]. The successful development of such reactions is an enormous challenge because of the need to balance an extra set of equilibria and competing side reactions for each new component of the reaction mixture. Most of the multicomponent reactions that have been developed so far have relied on the starting material's innate tendency to react, leaving researchers at the mercy of that compound's 'desires'. However, modern organic synthesis seeks to use engineered catalysts that override natural reactivity in an ordered fashion, methodically combining each component.

Decades of catalysis research have brought us to the point at which many kinds of molecule can be formed with a large excess of one of their mirror-image isomers (enantiomers)[5]. This has been a major focus of modern research, because often only one enantiomer of a drug will interact with a target enzyme. Catalysis offers an ideal solution to producing single enantiomers, because the source of enantioselectivity — a ligand at the reactive metal centre of the catalyst — is used in less than stoichiometric quantities. This greatly decreases costs. Although the development of such catalysts has been impressive, little progress has been made in combining multicomponent coupling reactions with enantioselective catalysis[3].

Meng *et al.* took this challenge head-on by developing a catalyst that promotes two enantioselective reactions between three compounds: a monosubstituted allene; bis(pinacolato) diboron ($B_2(pin)_2$); and an allylic phosphate (Fig. 1). Many possible products can form from this combination of reagents, and finding the perfect catalyst to tame the reactivity of all three starting materials is a daunting challenge.

On the basis of previous studies from the same laboratory[6], the authors knew that $B_2(pin)_2$ and the allene can react regioselectively (preferentially at a particular atom) and enantioselectively in the presence of a catalyst. However, adding an allylic phosphate to the reaction mixture created several problems for developing a two-step process. The catalyst must now differentiate between the double bonds in the allylic phosphate and the allene to perform the first transformation correctly. Furthermore, the second reaction — called an allylic substitution — must not only be enantioselective, but also yield linear, rather than branched products (for those in the know, it must react through an S_N2 rather than an S_N2' mechanism).

To achieve this, Meng *et al.* unsurprisingly had to perform extensive screening of catalyst ligands, and apply insight gained from experimental and computational mechanistic studies. This led to the discovery of a simple, copper-based catalyst that works for a wide range of substrates, and which provides high yields of product with the desired regio- and enantioselectivity.

One of the key advantages of the new multicomponent reaction is that the final products contain three functionalizable handles (groups that can be converted into a variety of other chemical motifs). The products should therefore be of great utility for synthesizing molecules for pharmaceutical or agricultural studies. To demonstrate this, Meng *et al.* completed two total syntheses of complex natural products using their reaction as the centrepiece. One of these syntheses — that of the antibiotic rottnestol — gave the natural product in almost seven times the overall yield of the next most efficient synthesis[7]. The authors prepared more than one gram of both compounds, highlighting the fact that their methodology can be reliably performed to generate meaningful amounts of material[8].

This work is a significant advance for transformations that combine enantioselective catalysis and multicomponent reactions, because it sets the powerful precedent of a single catalyst orchestrating the controlled union of many building blocks in precisely the right way. But as with any breakthrough reaction, there is still more to be done. The authors' substrates are only monosubstituted allenes (which have one group attached to the allene core) and disubstituted allylic phosphates (two groups attached to the allylic phosphate). Extension to more highly substituted starting materials should expand the number of accessible products. In the meantime, the startling discovery that a simple catalyst can tame the complex reactivity of this daunting set of molecules bodes well for future reactions in which bonds fall in line. ■

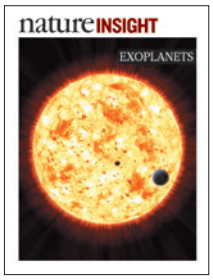
Matthew T. Villaume *and* **Phil S. Baran**
are in the Department of Chemistry, Scripps Research Institute, La Jolla, California 92037, USA.
e-mail: pbaran@scripps.edu

1. Meng, F., McGrath, K. P. & Hoveyda, A. H. *Nature* **513,** 367–374 (2014).
2. Keasling, J. D., Mendoza, A. & Baran, P. S. *Nature* **492,** 188–189 (2012).
3. de Graaff, C., Ruijter, E. & Orru, R. V. A. *Chem. Soc. Rev.* **41,** 3969–4009 (2012).
4. Joule, J. A. & Mills, K. *Heterocyclic Chemistry* 5th edn (Blackwell, 2010).
5. Noyori, R. *Asymmetric Catalysis in Organic Synthesis* (Wiley, 1994).
6. Meng, F., Jang, H., Jung, B. & Hoveyda, A. H. *Angew. Chem. Int. Edn* **52,** 5046–5051 (2013).
7. Czuba, I. R., Zammit, S. & Rizzacasa, M. A. *Org. Biomol. Chem.* **1,** 2044–2056 (2003).
8. Kuttruff, C. A., Eastgate, M. D. & Baran, P. S. *Nat. Prod. Rep.* **31,** 419–432 (2014).

nature INSIGHT

EXOPLANETS

18 September 2014 / Vol 513 / Issue No 7518



Cover illustration
Nik Spencer

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel: +44 (0) 20 7833 4000
e: nature@nature.com

nature publishing group

It is hard to imagine now, and the younger people in the field will not remember this, but there was a period when the search for exoplanets had rather a bad reputation, based on a number of high-profile claims that were subsequently disproved. Although there was broad agreement, even by the 1980s, that planet formation ought to be a natural part of the star-formation process, at least for low-mass stars, we were still basing our assumptions on what we might find using the Solar System as a template.

On a late-summer morning in 1995, I picked up a new manuscript by Michel Mayor and Didier Queloz. After reading the paper, I thought, 'this looks pretty promising'. My next thoughts were, how could a planet be so close to its parent star — it seemed very unlikely that it could form there — and was such a planet stable against evaporation by stellar radiation? I picked up the phone to a colleague knowledgeable in such things, and the second question was rapidly answered in the affirmative. The first question is still a topic of research, although the two main options of disk migration and gravitational scattering emerged quite rapidly.

Nineteen years later, this collection of exoplanet papers reviews the state of the field. It is only fitting that Mayor, along with his co-authors Christophe Lovis and Nuno Santos, provide an overview of where we stand today.

Most known exoplanets were discovered, through planetary transits of the parent star's disk, by the Kepler space telescope. Jack Lissauer, Rebekah Dawson and Scott Tremaine assess the highlights of the mission.

Adam Burrows goes on to look at our current theoretical understanding of exoplanets and their atmospheres.

The path for Kepler was blazed by the under-rated Convection, Rotation and Planetary Transits (CoRoT) mission, with some help from Microvariability and Oscillation of Stars (MOST). Artie Hatzes fills us in on what those missions found.

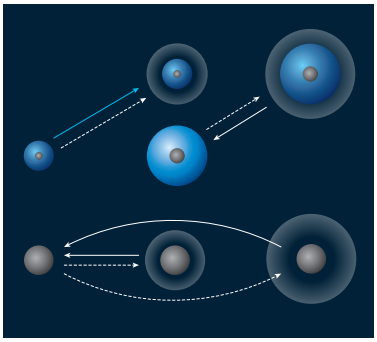
Finally, the precise instruments needed to measure the radial-velocity shifts of stars as they and their planets co-orbit the system's centre of mass, along with present and future instruments to better characterize the planetary atmospheres, are reviewed and anticipated by Francesco Pepe, David Ehrenreich and Michael Meyer.

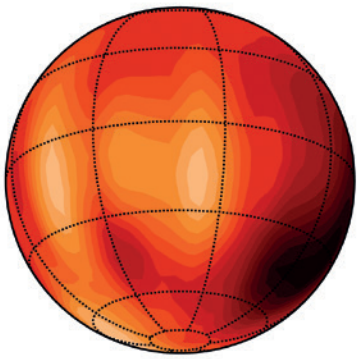
I hope that you enjoy this collection as much as I have enjoyed seeing history pass through my hands at *Nature*.

Leslie Sage
Astronomy Editor

CONTENTS

REVIEWS





REVIEW

doi:10.1038/nature13780

Doppler spectroscopy as a path to the detection of Earth–like planets

Michel Mayor[1], Christophe Lovis[1] & Nuno C. Santos[2,3]

Doppler spectroscopy was the first technique used to reveal the existence of extrasolar planetary systems hosted by solar–type stars. Radial–velocity surveys led to the detection of a rich population of super–Earths and Neptune–type planets. The numerous detected systems revealed a remarkable diversity. Combining Doppler measurements with photometric observations of planets transiting their host stars further provides access to the planet bulk density, a first step towards comparative exoplanetology. The development of new high–precision spectrographs and space–based facilities will ultimately lead us to characterize rocky planets in the habitable zone of our close stellar neighbours.

During the past three decades, the development of astronomical instrumentation and the scientific development of new observational techniques made it possible to transform the old philosophical concept of 'plurality of worlds' in the Universe into an active field of modern astrophysics. Today, almost 2,000 planets orbiting other stars are known, and we are contemplating an even more exciting challenge: discovering Earth-like exoplanets with physical conditions suitable for the complex chemistry of life to develop.

Some of the most important discoveries in this field have been made using the technique of Doppler spectroscopy. These results are the focus of this Review. They illustrate the tremendous progress that has been made in our understanding of exoplanet populations in the Galaxy, and the role of the stellar environment in the formation of planetary systems.

The discovery of a whole new population of planets orbiting other stars has now moved the focus of exoplanet researchers to two main areas: the search for planets of lower and lower mass, and the precise characterization of the new-found planets. In the years to come, the rise of a new set of experiments, including ground-based giant telescopes and space-based missions dedicated to the detection and characterization of planets hosted by bright stars, will allow the next big steps in this research. These efforts will bring us closer to the goal of detecting and characterizing Earth-like exoplanets of rocky composition orbiting within the habitable zone of their host star.

Early history

How many planets are there in the Milky Way? How many planets are similar to Earth? It is interesting to look at the astronomical literature of the twentieth century for estimations of the number of planetary systems in the Galaxy. Before 1943, the values ranged from zero to, at most, a few systems. The formation of protoplanetary gaseous nebulae was thought to result from the tidal capture of a stellar envelope through a close encounter with another star[1]. The extremely low probability of such a small impact collision was at the origin of these quite pessimistic estimations of number of planetary systems. In the early 1940s, claims of the discovery of several systems[2,3], later found to be false, induced, in a couple of years, a complete paradigm shift[4]. Those estimates jumped to billions if not hundreds of billions. It is interesting that such a drastic change of thought was the result of spurious detections of planetary systems.

The use of variation of stellar radial velocity due to gravitational interaction with a massive planet was suggested as a detection method long before spectrographs achieved the high precision needed for such detections[5,6]. The radial-velocity technique, based on the variable Doppler shift of stellar absorption lines, is able to measure planetary orbital period, orbital eccentricity and minimum mass ($M\sin i$). The amplitude of radial-velocity variations depends on the planet mass and orbital distance. In the Solar System, Jupiter induces a 12 m s^{-1} radial-velocity signal on the Sun with a periodicity of 12 years, whereas Earth imprints a tiny 0.1 m s^{-1} signal at a 1-year period. The corresponding Doppler shifts on the stellar spectrum are, however, extremely challenging to measure (~10^{-8}–10^{-10} of the wavelength), which hampered progress in this field for decades.

It was only during the 1980s that several ideas and technological solutions were proposed for new spectrographs, allowing radial-velocity precision of a couple of dozen metres per second[7]. Among the pioneers, credit has to be given to Campbell and Walker[7] for their survey of around 20 stars. With a hydrogen–fluoride absorption cell in front of their spectrograph, they demonstrated a radial-velocity precision of the order 15 m s^{-1}. However, at the end of many years of monitoring, their efforts obtained a negative result: no detection of Jupiter analogues orbiting their small stellar sample of solar-type stars[8]. Another survey was initiated by Marcy and Butler[9] in 1988 at the Lick Observatory. Their iodine absorption cell gave, at that time, a similar precision of about 15 m s^{-1}. The result, in 1994, of that survey was similar to the earlier one: no Jupiter analogues were found around 25 solar-type stars[9].

At the same time as these surveys of limited size, a few teams were operating efficient spectrographs of moderate precision (250–500 m s^{-1}) but on large stellar samples. Among the many thousands of stars surveyed (mostly the main sequence stars F, G, K and M), a few stars were used as standard by the different teams and provided dozens to hundreds of radial-velocity measurements. When analysing the velocities of one of these objects, HD 114762, Latham *et al.*[10] found a periodic variation of 84 days and an amplitude corresponding to a possible companion of 11 times the mass of Jupiter (M_J), on an eccentric orbit. Combining their data with complementary measurements acquired at Haute-Provence Observatory allowed the publication of a very precise orbit[10].

Was this companion a planet or low-mass brown dwarf? At that epoch, the community was inclined towards the second option — a result of its short period, rather large values for the orbital eccentricity and mass, all characteristics not expected for a gaseous giant planet similar to the ones of our Solar System. Based on the present observed diversity of detected exoplanets, this consensus is certainly

[1]Geneva Observatory, University of Geneva, 51 Chemin des Maillettes, 1290 Versoix, Switzerland. [2]Centro de Astrofisica e Departamento de Física e Astronomia, Faculdade de Ciências, Universidade do Porto, Rua das Estrelas, 4150-762 Porto, Portugal. [3]Instituto de Astrofísica e Ciências do Espaço, Centro de Astrofísica da Universidade do Porto, Rua das Estrelas, 4150-762 Porto, Portugal.

not a definitive conclusion. However, one characteristic should be mentioned: HD 114762 is a metal-deficient star (for which metallicity [Fe/H] = −0.7, where [Fe/H] = log[A_{Fe}/A_{H}] star − log(A_{Fe}/A_{H}) Sun and A is the abundance of a given chemical element). According to present-day observations and to state-of-the-art models of planetary formation, it seems difficult to form a massive planet in such a metal-poor environment[11] (see 'Chemical clues for stars with planets'). For instance, a recent high-precision 10-year-survey of more than 100 solar-type stars has not revealed one single gas-giant planet with metallicity significantly lower than −0.5 (ref. 12). In contrast with planet formation, the formation of low-mass stars is not strongly constrained by the metallicity of the star-formation environment. These facts suggest that HD 114762b is most likely to be a low-mass stellar companion. We should note, however, that a few low-mass companions with metallicity close to that of HD 114762 have been detected[11,13]. HD 114762 is the most massive of these outliers.

The discovery of 51 Pegasi b and its strange properties

At the beginning of the 1990s, two different approaches were used to determine precise stellar radial velocities. On the one hand, spectrographs with absorption cells in the beam of the spectrograph (hydrogen-fluoride cell or iodine cell)[14,15], and on the other hand, fibre-fed spectrographs with simultaneous calibration provided by a thorium lamp in a parallel fibre[16]. Both methods were aimed at providing a precise calibration in wavelength. In 1995, a comparable precision (15 m s^{-1}) was achieved by both techniques. However, one positive characteristic of the double-fibre spectrograph was its ability to obtain the final radial-velocity value a few seconds after the end of the observation sequence (an achievement not possible at the time for spectrographs using the absorption-cell technique). Furthermore, the double-fibre technique is more efficient in terms of photon noise, a crucial point to allow radial-velocity monitoring of a large sample of stars with moderate-sized telescopes.

In April 1994, with the new ELODIE spectrograph at Haute-Provence Observatory[16] (using the simultaneous calibration technique), Mayor and Queloz[17] initiated a systematic survey of 142 solar-type stars to search for brown dwarfs or giant planets. Included in that sample was 51 Pegasi, a metal-rich G2V type star, which was found to exhibit a periodic variation of its velocity with a period as short as 4.2 days. If resulting from the influence of a companion, the observed amplitude would indicate a minimum mass of a little less than half the mass of Jupiter[17]. This discovery revealed the first exoplanet hosted by a solar-type star and a first example of the family of so-called hot Jupiters.

Interestingly, such a short period was quite unexpected. The formation of gas-giant planets by agglomeration of ice particles was not supposed to be possible inside the 'ice-line'[18]. However, soon after this first detection, Lin *et al.*[19] showed that short-period gas-giant planets could result from the orbital migration of the young planet embedded in the accretion disk. This physical process was already described in the literature[20,21], but never incorporated in scenarios of planetary system formation.

Soon after the discovery of 51 Peg b, the detection of several short period planets was announced by Butler *et al*[22]. Clearly, 51 Peg b was not a unique object with exceptional characteristics.

Despite the run of detections, not all the community was convinced that these unexpected objects with short periods were indeed planets. However, a few crucial observations played a significant part in confirming their planetary nature. The detection of multi-planetary systems was strong evidence supporting the planetary explanation, but the most important observation was the first detection of a planetary transit.

HD 209458b is a hot Jupiter-like planet with an orbital period close to 3.5 days. During the night of the 9 September 1999, at the precise time derived from the radial-velocity ephemerides, the first transit of the planet was detected, this was followed by a second detection 7 days later[23]. The same host star was also scrutinized by another team and the transit detected[24]. The data also allowed researchers to derive the mean density of that gas giant, showing that it was as low as 0.3 g cm^{-3}, less than half the mean density of Saturn. Observation of the transit was repeated with the Hubble Space Telescope the following year[25]. The amazing precision of that transit is a milestone of exoplanet research. Hot Jupiters are indeed real gas-giant planets. Not only did the transit curve put an end to alternative interpretations for the existence of hot Jupiters, but that observation, with its remarkable precision, also opened the door to space experiments dedicated to exoplanetary transits such as the Convection, Rotation and Planetary Transits (CoRoT) and Kepler missions, and future missions such as Transiting Exoplanet Survey Satellite (TESS), Characterising Exoplanet Satellite (CHEOPS) and Planetary Transits and Oscillations of Stars (PLATO).

A recent result refined our knowledge of 51 Peg b. Observing high-resolution spectra of stars hosting planets, it is possible to detect spectral fingerprints of a planet's atmosphere. Using this technique, significant absorption from carbon monoxide and water vapour were observed in the dayside atmosphere of 51 Peg b[26]. In this way, the radial velocity of the planet could be measured directly, allowing the determination of the planet/star mass ratio and orbital inclination. This gave a direct estimate of the mass of 51 Peg b of 0.46 M_J.

An explosion of discoveries

In 1995, the radial-velocity precision achieved by the best instruments was about 15 m s^{-1}. By 1996, improvements in the data-reduction software allowed the precision of the iodine-cell technique to be improved down to 3 m s^{-1} using the Hamilton Echelle Spectrometer at the Lick Observatory and Keck High Resolution Echelle Spectrometer (HIRES)[27]. The need to increase the precision of Doppler measurements is obvious because the amplitude of the radial-velocity wobble is directly proportional to the planetary mass.

Soon after the discovery of 51 Peg b[17], existing radial-velocity surveys of nearby solar-type stars were significantly expanded, and new ones started, with precision of 3–10 m s^{-1} (refs 16, 27–31). Additional hot-Jupiter discoveries quickly followed, aided by the relative ease of detection of their radial-velocity signals[22]. Over the next two decades, several hundred giant exoplanets were found, spanning a wide range of mass and orbital distance.

In 2003, a new gain in precision was achieved with the construction

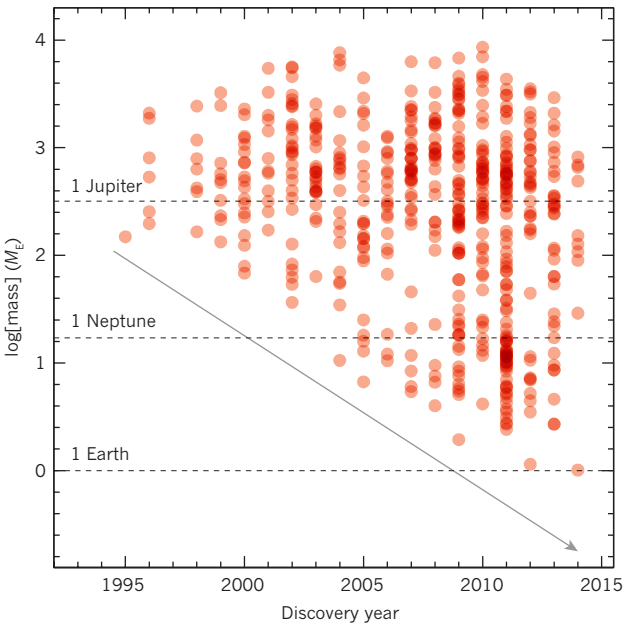


Figure 1 | Exoplanet discoveries as a function of time. The plot shows the minimum mass of the planets discovered by radial-velocity surveys as a function of discovery epoch. The horizontal lines denote the position of Earth, Neptune and Jupiter in this plot. The lower envelope of the points is illustrated by the solid line. This plot shows the incredible decrease in mass of the discovered planets, reflecting the increasing precision of radial-velocity surveys. Earth-mass planets are presently within reach and have been detected.

Table 1 | Some remarkable planetary systems discovered or characterized with Doppler spectroscopy

System name	Description	Comments
51 Pegasi[17]	1 hot Jupiter	First exoplanet found around a solar-type star
υ Andromedae[101]	3 gas giants within 2.5 AU	First multi-planet system identified
HD 209458 (refs 23, 24, 102)	1 transiting hot Jupiter	First transiting exoplanet found and well suited to atmospheric characterization owing to host-star brightness
HD 80606 (refs 103, 104)	1 transiting hot Jupiter	Highest known orbital eccentricity (e = 0.93) and misaligned orbit
GJ 436 (refs 57, 73)	1 transiting Neptune	First transiting Neptune, close-in but eccentric orbit and orbiting a nearby M dwarf
μ Arae (refs 56, 105)	1 close-in Neptune, 3 gas giants within 5 AU	Dynamically packed system of giant planets with inner low-mass object
55 Cnc[58,106]	1 transiting super-Earth, 4 gas giants within 6 AU	Dynamically packed system of giant planets with inner low-mass and intermediate-density object
HD 189733 (ref. 107)	1 transiting hot Jupiter	Well suited to atmospheric characterization owing to host-star brightness
HD 149026 (ref. 108)	1 transiting hot Saturn	Dense giant planet with large heavy element core
HD 69830 (ref. 59)	3 Neptunes within 0.6 AU	First system of close-in, low-mass planets
GJ 581 (refs 109, 110)	At least 2 super-Earths and 1 Neptune	First compact, low-mass system around an M dwarf
XO-3 (ref. 111)	1 transiting super-Jupiter	First planet showing a spin-orbit misalignment
HD 45364 (ref. 39)	2 gas giants within 0.9 AU	System in 3:2 mean motion resonance
HD 40307 (refs 79, 112)	4 super-Earths within 0.6 AU	Compact low-mass system with a potentially habitable planet
GJ 1214 (ref. 71)	1 transiting mini-Neptune	Low-mass, low-density object orbiting a nearby late M star
GJ 876 (ref. 113)	1 close-in super-Earth, 2 gas giants and 1 Neptune within 0.33 AU	Three outer planets locked in a 4:2:1 Laplace resonance similar to the Galilean moons of Jupiter
WASP-8 (ref. 114)	1 transiting hot Jupiter	Retrograde and misaligned orbit
HD 10180 (ref. 115)	Up to 7 planets within 3.5 AU, mostly Neptunes	Most populated exoplanet system known so far
HD 85512 (ref. 76)	1 super-Earth at 0.26 AU	Potentially habitable planet with a radial-velocity amplitude of 0.7 m s^{-1}
HD 97658 (ref. 69)	1 transiting super-Earth	Intermediate-density object orbiting a bright K dwarf
GJ 3470 (ref. 72)	1 transiting Neptune	Nearby M-dwarf host
α Centauri B[33]	1 short-period Earth-mass planet	Closest planetary system to the Sun
GJ 667C[78]	2 super-Earths	Potentially habitable planet

of the High Accuracy Radial Velocity Planet Searcher (HARPS) spectrograph at La Silla Observatory in Chile[32]. This fibre-fed vacuum spectrograph allows routine precision better than 1 m s^{-1}. This is still the most precise instrument for exoplanet detection. Following this and other developments, a large number of systems with planets smaller than Neptune could be detected (Fig. 1). Especially striking is the continuous decrease in the mass of detected exoplanets — an amazing improvement from the 3,000 M_E (where M_E is the mass of Earth) of HD 114762 in 1989, to the 150 M_E of 51 Peg in 1995, down to 1.1 M_E for the companion of α Centauri B in 2012 (ref. 33).

Ensemble properties of exoplanets

After the initial discovery phase, it became possible to derive unbiased exoplanet population statistics by quantifying detection efficiencies as a function of planet parameters (orbital period, mass and eccentricity). In this Review, we summarize the results of various attempts to characterize the ensemble properties of exoplanets using the radial-velocity technique. We complement these with an overview of the transit searches for hot Jupiters; these have also developed tremendously over the past decade. We restrict ourselves to ground-based surveys (see the Review by Lissauer *et al.* on page 336 for results of Kepler mission).

Statistics of gas-giant planets

Soon after the discovery of 51 Peg b, it became clear that short-period gas giants are relatively rare. Globally, early radial-velocity surveys mainly revealed a population of gas giants at orbital distances of 1–5 AU[34] (1 AU is the Sun–Earth distance). Key characteristics of this population include[35,36] an overall occurrence rate of about 15% (minimum mass $M\sin i > 50\ M_E$, orbital period (P) < 10 years); a mass distribution peaking at ~1–2 M_J with a 'brown dwarf desert' above 10–20 M_J; a wide distribution of orbital eccentricities that differs markedly from the low eccentricities seen in the Solar System; a higher metallicity of the host stars when compared with the average value found for their field star counterparts; and an overall occurrence rate of 1.05 ± 0.26% for hot Jupiters[35,37], valid for planets with $M\sin i > 0.1\ M_J$ and $P < 10$ days.

We note that our knowledge of the period distribution is at present limited by the duration of radial-velocity surveys and the sampling of long-period signals. We stress the importance of continuing these programmes for at least a decade to thoroughly probe the 5–10 AU region of planetary systems, where the formation of giant planets is likely to be most efficient. This is also the region where significant overlaps with direct imaging and microlensing techniques are expected.

In many cases, not one but several giant planets have been found in the same system[34,38]. Various types of dynamical configurations exist, from widely separated orbits to strongly interacting mean motion resonances[39]. Owing to the compactness and proximity of such resonances, the dynamical stability of several systems is not *a priori* obvious and must to be probed by dedicated numerical integrations. In favourable cases, planet–planet interactions are sufficiently strong to affect radial velocities in a measurable way (and on orbital timescales), which then yields direct constraints on the inclination angles of the orbits and true masses of the planets[40]. There are some notable examples illustrating the diversity of the population of giant planets (Table 1).

So far, perhaps the most striking result concerning long-period giant planets has been their tendency to have high orbital eccentricities (median value of about 0.3). The standard scenario of planet formation within a protoplanetary disk calls for orbits to be much closer to circular. The Solar System has long been seen as a prototypical example of this model. Clearly, the formation and evolution of planetary systems is generally much more complex than originally thought.

Strong gravitational interactions between giant planets after disk dissipation[41,42], as well as the gravitational influence of bound or passing stellar companions[43], probably have a crucial role in the evolution and the final shaping of planetary systems. In this context, a major question that has yet to be answered is how common Solar System analogues are; that is, systems whose dynamics are dominated by a massive gas giant on a low-eccentricity orbit at several astronomical units from the star.

Insights from ground-based transit searches

Soon after the discovery of the first transiting giant planet, HD 209458b[23,24], several ground-based efforts started to target the population of hot Jupiters through the photometric transit technique. Exoplanet transits mainly provide planetary orbital period, inclination and planet size (radius). Coupled with radial-velocity measurements, which provide the planetary mass, these can be used to derive the planet bulk density. This has been the main observational method to constrain the internal structure of exoplanets used so far.

Although early transit-search attempts were plagued by insufficient precision and inefficient operations, more recent large-scale surveys such as Wide Angle Search for Planets (WASP)[44] and Hungarian Automated Telescope Network (HATNet)[45] have eventually unveiled hundreds of transiting hot Jupiters orbiting FGK dwarfs within about 200 pc of the Sun. Two key properties of this population are a planet-size distribution that shows an excess of anomalously large radii, hinting at an as yet poorly understood physical mechanism that injects or traps excess internal heat inside the planet[46–50]; and the existence of a sub-population of hot Jupiters whose orbital plane is misaligned with respect to the stellar equatorial plane[51] (Fig. 2), preferentially found around hotter stars (with effective temperature ($T_{\rm eff}$) of more than 6,250 K)[52].

The formation and evolution of hot-Jupiter systems is a matter of active research and a full picture is still missing. For a long time the canonical scenario of inward migration within a protoplanetary disk prevailed, but the discovery of misaligned hot Jupiters has significantly changed this. It now seems clear that dynamical interactions between multiple giant planets, and/or interactions with massive outer companions (for example, Kozai oscillations), have a major role during or after planet formation[53–55]. This echoes the conclusions already drawn from the observed high eccentricities of longer-period giant planets. In those scenarios, planets are perturbed into orbits with high eccentricity and potentially high inclination. These are then circularized and realigned with the star through tidal interactions between the planet and the star. Indeed, the convergence between observed spin-orbit alignment and the existence of an outer stellar convection zone for cool stars ($T_{\rm eff}$ < 6,250 K) points towards the influence of tides on the orbital evolution of at least some of the known hot Jupiters.

The rise of Neptunes and super-Earths

Improvements in radial-velocity precision towards 1 m s^{-1}, coupled with dedicated observing strategies, opened a new search space for radial-velocity surveys. In 2004, three planets in the Neptune-mass range were found for the first time: μ Arae c, GJ 436b and 55 Cnc e[56–58]. In 2006, HD 69830 was found to be the first system of multiple low-mass planets on close-in orbits, an architecture that would later prove to be common[59]. By 2008, it had become clear from the HARPS survey[32] that a large population of Neptunes and super-Earths exists on short-period orbits[60], with a preliminary occurrence rate of 30% for such planets ($M\sin i$ < 30 $M_{\rm E}$, P < 50 days). In 2011, early results from the space-based Kepler mission fully confirmed this picture and greatly expanded the landscape of small-planet studies[61]. (The Kepler results are discussed in the Review by Lissauer *et al.* on page 336.)

Only three radial-velocity programmes discovered a sufficient number of low-mass planets ($M\sin i$ < 30 $M_{\rm E}$) to allow meaningful statistical studies of this population. These are the HARPS survey for FGK stars (44 detections[32]), the HARPS survey of M dwarfs (10 detections[62]), and the Keck-HIRES survey of FGK stars (9 detections[63]). Key properties of the low-mass population unveiled by the HARPS surveys can

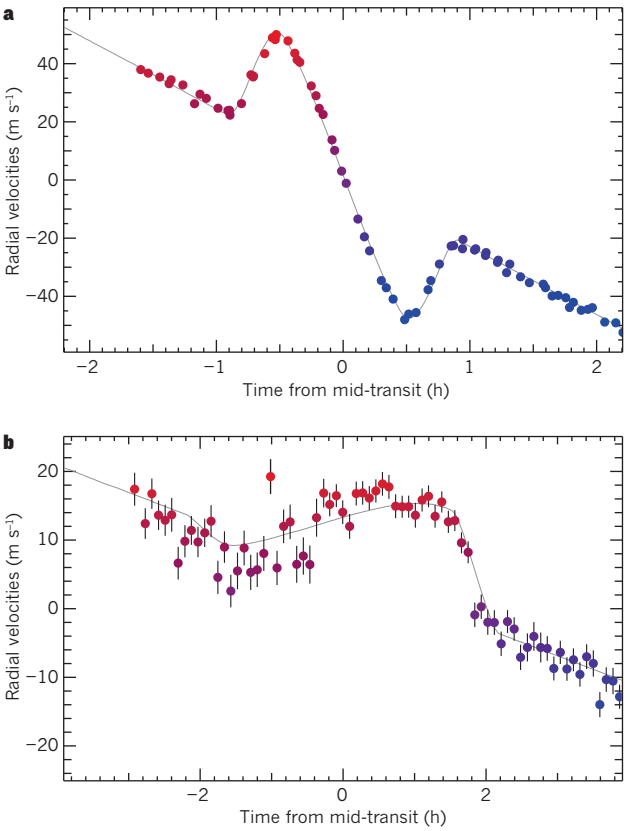


Figure 2 | Rossiter–McLaughlin effect for probing the spin-orbit alignment of exoplanets. For transiting exoplanets, precise radial-velocity measurements are not only able to measure exoplanet mass, but also the sky-projected angle between the planet's orbital plane and the stellar equatorial plane by the Rossiter–McLaughlin (RM) effect. The technique relies on the selective occultation of approaching and receeding parts of the rotating stellar disk by the planet during transit, causing variable net Doppler shifts that are detectable as deformations of the stellar spectral lines. The RM effect has a duration equal to the transit duration (typically hours) and is superimposed onto the usually larger radial-velocity signal caused by the gravitational pull of the planet on its star, which has periodicity equal to the planet orbital period (typically days for hot Jupiters). Two examples obtained with the HARPS spectrograph are shown. **a**, The exoplanet HD 189733A b exhibits a symmetric RM effect with a positive-then-negative radial-velocity anomaly that is the signature of an aligned and prograde star–planet system[116]. **b**, The WASP-8A system, however, shows a clearly asymmetric RM effect caused by a planet on a retrograde and strongly misaligned orbit with respect to the stellar rotation axis[114].

be summarized as follows: for FGK stars, a global occurrence rate of 0.33 ± 0.05 planets per star for $M\sin i$ between 3 $M_{\rm E}$ and 30 $M_{\rm E}$ and P < 50 days; a mass distribution showing a steep rise below 20 $M_{\rm E}$, with a continuously rising trend at least down to 5 $M_{\rm E}$; a multiplicity rate of at least 70% among systems with at least one Neptune or super-Earth, that is, most low-mass planets are found in multi-planet systems; for M dwarfs, a global occurrence rate of 0.40 ± 0.11 planet per star for $M\sin i$ between 3 $M_{\rm E}$ and 30 $M_{\rm E}$ and P < 50 days; and an occurrence rate of super-Earths (1–10 $M_{\rm E}$) in the habitable zone of M dwarfs of $0.41^{+0.54}_{-0.13}$ planets per star.

The high occurrence rates have one key consequence: systems of multiple planets with masses between 1 $M_{\rm E}$ and 20 $M_{\rm E}$ orbiting within 0.5–1.0 AU represent the most common type of planetary systems in the Galaxy. This result is supported by planet population synthesis models[11]. Moreover, such systems often exhibit a packed dynamical configuration, with little space left for stable orbits between consecutive planets (see Table 1 for examples of such compact systems). Planet formation and evolution scenarios must now account for the existence of this

population. Competing theories include convergent migration of several protoplanets within a disk towards the inner regions of the system[64], and *in situ* formation of super-Earths or Neptunes within a disk that is significantly more massive than the Minimum Mass Solar Nebula[65].

Essential clues on the formation path of these planets will come from the study of their internal structure. A key question is whether volatiles (mainly water) are a significant constituent of the interiors, which would point to a formation beyond the ice line. Another open question is to what extent the prevalence of H/He envelopes is a function of planet mass, formation path and irradiation. These issues can be investigated by discovering super-Earths and Neptunes transiting nearby bright stars; a precise mass and radius can be obtained for these planets and atmospheric composition can be studied. At present, there are six such planets (55 Cnc e[58,66–68], HD 97658b[69,70], GJ 1214b[71], GJ 3470b[72], GJ 436b[57,73], and HAT-P-11b[74]) with radius (R) < 6 R_E, all of them discovered from the ground. The space missions CoRoT and Kepler have also provided a sample of objects with precisely-measured densities (see Review by Lissauer *et al.* on page 336). However, these targets are generally much more distant than those discovered by radial-velocity surveys, and therefore more difficult to characterize.

Although still limited, the sample of low-mass planets with well-measured densities already shows a wide diversity of compositions (Fig. 3 gives an overview of our present knowledge of Neptunes and super-Earth mean densities; see ref. 75 for densities obtained from the Kepler results and radial-velocity follow-up).

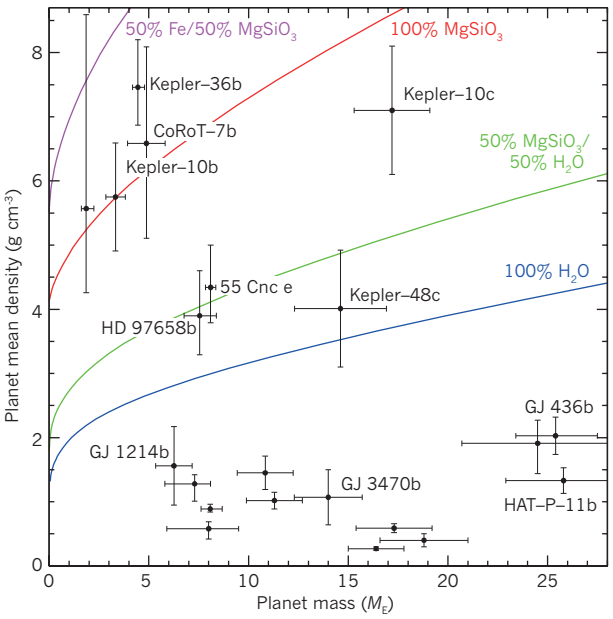


Figure 3 | Mass–density diagram for Neptunes and super-Earths. Few low-mass exoplanets have precise mass and radius measurements from which a reliable density can be derived. Here we show the mass and density of the 21 Neptunes and super-Earths that have a mass measurement with better than 20% precision. A population of low-density objects can be seen below around 2.0 g cm^{-3}, indicating a substantial H/He envelope much like Uranus and Neptune. Another population of much denser objects is also revealed, indicating bulk compositions ranging from terrestrial to more volatile-rich (for example, H_2O). The overall trend indicates lower densities towards higher masses. However, high-density objects seem to exist also at masses above 10 M_E, while low-density mini-Neptunes occur at masses of only a few Earth masses (M_E). The planets GJ 1214b, HD 97658b and 55 Cnc e span a narrow mass range of 6–8 M_E and have mean densities from 1.6 g cm^{-3} (GJ 1214b) to almost 5 g cm^{-3} (55 Cnc e), indicating very different internal structures at a given mass. This hints at a complex mass–radius relationship for low-mass exoplanets that does not depend on mass (or radius) alone, but also on environmental effects related to the formation and evolution of planetary systems. Mass–density relations from internal structure models with various bulk compositions are superimposed onto the observations[117]. For clarity, only a selection of the exoplanets are labelled.

Finally, radial-velocity surveys have recently come tantalizingly close to discovering planets in the habitable zone. The GJ 581 and HD 85512 systems comprise at least one super-Earth that could lie at the edge of habitability, depending on surface conditions[76,77]. Moreover, the planets GJ 667C c and HD 40307g are located within the classical habitable zone[78–80]. However, because neither the bulk density nor the atmospheric composition of these worlds is known, all discussions about their habitability remain largely speculative.

The discovery of such objects has been possible thanks to sub-metre-per-second radial-velocity precision and a careful analysis of the radial-velocity time series. At this level of precision, however, various physical phenomena in stellar photospheres contribute significant signals that can hide planetary radial-velocity signatures if not properly modelled. Solar-type stars have an outer convective envelope that exhibits variability on different timescales. Granulation, magnetic features (such as cool spots, plages and faculae) and long-term activity cycles all induce radial-velocity variability at the metre-per-second level[81,82]. Understanding how to diagnose and correct these effects is an active area of research[83–85]. State-of-the-art instrumentation and dense temporal sampling are the key to making progress in this field, and to ultimately push radial-velocity sensitivity down to the 10 cm s^{-1} level. This is the realm of habitable, Earth-mass planets around solar-type stars. Considering that HARPS has already detected planetary signals with an amplitude of 50 cm s^{-1} (ref. 32), and that modelling of stellar signals is still in its infancy, the exploration of the habitable zone around nearby FGK and M dwarfs is within reach of the radial-velocity technique. That is the main goal of future Doppler instruments (for example, the Echelle Spectrograph for Rocky Exoplanet and Stable Spectroscopic Observations (ESPRESSO) on the European Southern Observatory Very Large Telescope (ESO-VLT), see Review by Pepe *et al.* on page 358).

Chemical clues for stars with planets

One of the most crucial pieces of information to understand the properties of the discovered planets and to access their formation process comes from the study of planet host stars. On the one hand, precise stellar parameters, such as the radius, are crucial if we want to measure precise values for the radius of a transiting planet[86]. On the other hand, the chemical composition of a planet, both its interior and atmosphere, is also likely to be related to the chemical composition of the protostellar cloud, reflected in the composition of the stellar atmosphere[87]. The precise derivation of stellar chemical abundances thus provides important clues to understanding the planets and their observed properties.

Furthermore, a number of studies have pointed towards the existence of a strong relationship between the properties and frequency of the new-found planets and those of their host stars. Large spectroscopic studies[88,89] confirmed initial suspicions of a positive correlation between the probability of finding a giant planet and the metal content of the stars (Fig. 4). Curiously, this strong metallicity–giant-planet correlation was not found for the lowest mass planets[90,91].

It was soon realized that this correlation for giant planets was a key aspect of understanding planet formation. The simple existence of such a correlation has been pointed out as a strong evidence for giant-planet formation through the core-accretion process[11]. The lack of correlation for the lower-mass planets is also in full agreement with the expectations from such models. Indeed, stars formed out of metal-rich clouds are expected to have a higher mass of solid elements in their protoplanetary disks, thus leading to the formation of planet cores over a short timescale. These are then able to accrete gas and become giant planets. However, stars formed out of metal-poor clouds will not have much planet-forming material in their disks. The planets will grow slowly, never achieving enough mass to become giant planets.

Recent results from a specific survey for giant planets orbiting a sample of metal-poor stars was conducted with the Doppler velocimetry technique, using the HARPS spectrograph[12]. The results fully confirm the lower frequency of giant planets orbiting lower metallicity stars,

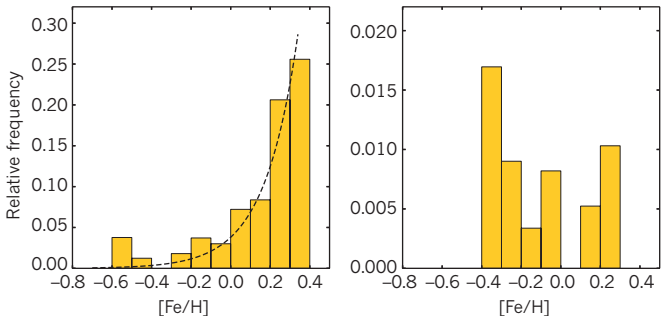


Figure 4 | Metallicity distribution of planet-hosting stars. In the left panel, the frequency of giant planets as a function of stellar metallicity is shown based on results from the HARPS planet search programme. The dashed line shows a power-law fit to the histogram values. In the right panel, we present the same plot but for stars that host only Neptune- or super-Earth-like planets. By metallicity we denote the abundance (A) of iron relative to the Sun, [Fe/H] = $\log(A_{Fe}/A_H)$ star – $\log(A_{Fe}/A_H)$ Sun. These plots show a clear correlation between the presence of giant planets and the metallicity of the star. This trend is not seen for stars hosting lower-mass planets (as in ref. 90).

and point to a possible limit in metallicity, below which no giant planets may be formed.

Further evidence for the planet-formation process comes from the study of specific elemental abundances. It is now becoming clear that the abundance of α elements plays an important part in the formation of planetary systems, particularly in metal-poor environments[92]. The role of the abundances of other elements is also under discussion; some curious trends are a matter of strong debate, for example the abundances of the light element lithium[93,94], its isotope lithium-6 (refs 95, 96) or other elements[97,98].

It is important to note that the role of stellar metallicity in the formation of different architectures of planetary systems has also been addressed. Recently, suspicions have been raised concerning the influence of stellar metallicity on the orbital period of planets[99,100] — planets orbiting metal-poor stars have longer periods than those in metal-rich systems. These results show that metallicity is one of the most crucial ingredients in the formation of planetary systems, controlling not only the planet-formation efficiency, but also the outcome of the planet-formation process, including mass, composition and architecture.

Future prospects

How will the field be moved forward? Almost 20 years after the discovery of 51 Peg b, the field of exoplanets has reached maturity, but our knowledge remains patchy and exoplanet parameter space has not been explored systematically. Clearly, the future lies in the detection and full characterization of entire planetary systems around nearby bright stars, for which precision measurements of both the stellar and planetary parameters can be obtained. To this purpose, a wealth of ground-based and space-based facilities will be working together in a common effort.

The search for transits of super-Earths and Neptunes around bright stars will be carried out by the Next Generation Transit Search (NGTS, in 2014), the ongoing MEarth project and other ground-based surveys; and by the TESS (in 2017), Kepler-K2 (in 2014) and PLATO (by 2024) missions from space. PLATO in particular will detect transiting Earth-like planets in the habitable zone of nearby solar-type stars. At the same time, radial-velocity surveys using spectrographs such as HARPS, HARPS-N, Keck HIRES, Automated Planet Finder (APF) at Lick, ESPRESSO (to come into use in 2016), CARMENES (to begin work in 2015) and SPiROU (to begin in 2017) will continue the thorough exploration of planetary systems in the solar neighbourhood, and will carry out the follow-up of the above-mentioned transit search missions to measure planet masses. The CHEOPS mission (in 2017) will provide essential support to both radial-velocity and transit surveys through a flexible high-precision photometric follow-up from space. In addition, the European Space Agency's ongoing Gaia mission will provide high-accuracy fundamental stellar parameters for all planet host stars, and, coupled to high-resolution spectroscopy from the ground, will greatly improve the achievable precision of planetary masses and radii. Gaia will also detect giant planets at intermediate semi-major axes, complementing radial-velocity surveys and high-contrast imaging in an effort to fully explore planetary systems, including dynamically important gas giants on long-period orbits.

By between 2020 and 2025 the exoplanet landscape will, therefore, offer a tantalizing collection of objects spanning the whole parameter space and including terrestrial planets with habitable surface conditions. There is a clear path forward for finding the 'best' such planets — those that are closest to the Sun and most amenable to further characterization. Not only will their bulk composition be well constrained, but also their atmospheres will be probed with techniques such as transmission spectroscopy (primary transit) and emission spectroscopy (secondary eclipse). The James Webb Space Telescope (launching in 2018) will lead these efforts, complemented by high-resolution spectroscopy from the ground with the future extremely large telescopes. We are lucky enough to live in a time in which humans are, for the first time, contemplating the realistic possibility of exploring other planets similar to our own. Whether they will be few or plenty, orbiting a Sun-like star or an M dwarf, rocky or ocean worlds, with Earth-like atmospheres or more exotic ones, remains to be seen. This makes the quest all the more exciting. ■

Received 30 April; accepted 24 July 2014.

1. Jeans, J. *Problems of Cosmogony and Stellar Dynamics*, p. 290 (Cambridge University Press, 1919).
2. Strand, K. 61 Cygni as a triple system. *Publ. Astron. Soc. Pacif.* **55,** 29–32 (1943).
3. Reuyl, D. & Holmberg, E. On the existence of a third component in the system 70 ophiuchi. *Astrophys. J.* **97,** 41 (1943).
4. Dick, S. J. in *Bioastronomy – The Search for Extraterrestrial Life* (eds J. Heidmann & M.J. Klein) 356–363 (Springer, 1991).
5. Belorizky, D. Le Soleil, étoile variable. *L'Astronomie* **52,** 359–361 (1938).
6. Struve, O. Proposal for a project of high-precision stellar radial velocity work. *Observatory* **72,** 199–200 (1952).
7. Campbell, B. & Walker, G. A. H. in *Stellar Radial Velocities: IAU Colloquium* 88 (eds Davis Philip, A.G. & Latham, D.) 5–18 (L. Davis, 1985).
8. Walker, G. A. H. *et al.* A search for Jupiter-mass companions to nearby stars. *Icarus* **116,** 359–375 (1995).
9. Marcy, G. W. & Butler, R. P. in *The Bottom of the Main Sequence and Beyond* (ed. Tinney, C.) 98 (Springer, 1994).
10. Latham, D. W., Stefanik, R. P., Mazeh, T., Mayor, M. & Burki, G. The unseen companion of HD114762 – a probable brown dwarf. *Nature* **339,** 38–40 (1989).
11. Mordasini, C., Alibert, Y., Benz, W., Klahr, H. & Henning, T. Extrasolar planet population synthesis. IV. Correlations with disk metallicity, mass, and lifetime. *Astron. Astrophys.* **541,** A97–A119 (2012).
Exoplanet population synthesis models, which try to reproduce ensemble properties of exoplanets from protoplanetary disk evolution.
12. Santos, N. *et al.* The HARPS search for southern extrasolar planets. XXV. Results from a metal-poor sample. *Astron. Astrophys.* **526,** A112–A128 (2011).
13. Santos, N. *et al.* SWEET-Cat: a catalogue of parameters for stars with exoplanets. I. New atmospheric parameters and masses for 48 stars with planets. *Astron. Astrophys.* **556,** A150 (2013).
14. Campbell, B., Walker, G. A. & Yang, S. A search for substellar companions to solar-type stars. *Astrophys. J.* **331,** 902–921 (1988).
15. Marcy, G. W. & Butler, R. P. Precision radial velocities with an iodine absorption cell. *Publ. Astron. Soc. Pacif.* **104,** 270–277 (1992).
16. Baranne, A. *et al.* ELODIE: a spectrograph for accurate radial velocity measurements. *Astron. Astrophys.* **119,** 373–390 (1996).
17. Mayor, M. & Queloz, D. A Jupiter-mass companion to a solar-type star. *Nature* **378,** 355–359 (1995).
This paper reports the discovery of 51 Peg b, the first confirmed exoplanet around a solar-type star.
18. Boss, A. Proximity of Jupiter-like planets to low-mass stars. *Science* **267,** 360–362 (1995).
19. Lin, D. N. C., Bodenheimer, P. & Richardson, D. C. Orbital migration of the planetary companion of 51 Pegasi to its present location. *Nature* **380,** 606–607 (1996).
This was the first attempt at explaining the existence of hot Jupiters in terms of orbital migration within a protoplanetary disk.
20. Goldreich, P. & Tremaine, S. Disk-satellite interactions. *Astrophys. J.* **241,** 425–441 (1980).

21. Lin, D. N. C. & Papaloizou, J. On the tidal interaction between protoplanets and the protoplanetary disk. III – orbital migration of protoplanets. *Astrophys. J.* **309,** 846–857 (1986).
22. Butler, R. P., Marcy, G. W., Williams, E., Hauser, A. & Shirts, P. Three new 51 Pegasi-type planets. *Astrophys. J.* **474,** L115–L118 (1997).
This article provided confirmation that 51 Peg b is not unique: hot Jupiters exist around many stars.
23. Charbonneau, D., Brown, T. M., Latham, D. W. & Mayor, M. Detection of planetary transits across a sun-like star. *Astrophys. J.* **529,** L45–L48 (2000).
This article reports the first detection of an exoplanetary transit.
24. Henry, G. W., Marcy, G. W., Butler, R. P. & Vogt, S. S. A transiting 51 Peg-like planet. *Astrophys. J.* **529,** L41–L44 (2000).
25. Brown, T. M., Charbonneau, D., Gillilland, R. L., Noyes, R. W. & Burrows, A. Hubble space telescope time-series photometry of the transiting planet of HD 209458. *Astrophys. J.* **552,** 699–709 (2001).
26. Brogi, M. *et al*. Detection of molecular absorption in the dayside of exoplanet 51 Pegasi b. *Astrophys. J.* **767,** 27–36 (2013).
27. Butler, R. P. *et al*. Attaining Doppler precision of 3 m/s. *Publ. Astron. Soc. Pacif.* **108,** 500–509 (1996).
28. Cochran, W. D. & Hatzes, A. P. in *Precise Stellar Radial Velocities IAU Colloquium 170* (eds Hearnshaw, J. B. & Scarfe, C. D.) 113 (Astron. Soc. Pacif., 1999).
29. Queloz, D. *et al*. The CORALIE survey for southern extra-solar planets. I. A planet orbiting the star Gliese 86. *Astron. Astrophys.* **354,** 99–102 (2000).
30. Endl, M., Kürster, M. & Els, S. The planet search program at the ESO Coudé Echelle spectrometer. I. Data modeling technique and radial velocity precision tests. *Astron. Astrophys.* **362,** 585–594 (2000).
31. Tinney, C. G. *et al*. First results from the Anglo-Australian planet search: a brown dwarf candidate and a 51 Peg-like planet. *Astrophys. J.* **551,** 507–511 (2001).
32. Mayor, M. *et al*. Setting new standards with HARPS. *Messenger* **114,** 20–24 (2003).
33. Dumusque, X. *et al*. An Earth-mass planet orbiting α Cen B. *Nature* **491,** 207–211 (2012).
This article reports the discovery of an Earth-mass planet on a short-period orbit around α Cen B, our closest stellar neighbour.
34. Udry, S. & Santos, N. C. Statistical properties of exoplanets. *Annu. Rev. Astron. Astrophys.* **45,** 397–439 (2007).
35. Mayor, M. *et al*. The HARPS search for southern extra-solar planets. Occurrence, mass distribution and orbital properties of super-Earths and Neptune-mass planets. Preprint at: http://arxiv.org/abs/1109.2497 (2011)
36. Cumming, A. *et al*. The Keck planet search: detectability and the minimum mass and orbital period distribution of extrasolar planets. *Publ. Astron. Soc. Pacif.* **120,** 531–554 (2008).
37. Wright, J. T. *et al*. The frequency of hot Jupiters orbiting nearby solar-type stars. *Astrophys. J.* **753,** 160–164 (2012).
38. Wright, J. T. *et al*. Ten new and updated multiplanet systems and a survey of exoplanetary Systems. *Astrophys. J.* **693,** 1084–1099 (2009).
39. Correia, A. C. M. *et al*. The HARPS search for southern extra-solar planets. XVI. HD 45364, a pair of planets in a 3:2 mean motion resonance. *Astron. Astrophys.* **496,** 521–526 (2009).
40. Correia, A. C. M. *et al*. The HARPS search for southern extra-solar planets. XIX. Characterization and dynamics of the GJ 876 planetary system. *Astron. Astrophys.* **511,** A21 (2010).
41. Chatterjee, S., Ford, E. B., Matsumura, S. & Rasio, F. A. Dynamical outcomes of planet-planet scattering. *Astrophys. J.* **686,** 580–602 (2008).
42. Ford, E. B. & Rasio, F. A. Origins of eccentric extrasolar planets: testing the planet-planet scattering model. *Astrophys. J.* **686,** 621–636 (2008).
43. Takeda, G. & Rasio, F. A. High orbital eccentricities of extrasolar planets induced by the Kozai mechanism. *Astrophys. J.* **627,** 1001–1010 (2005).
44. Pollacco, D. L. *et al.* The WASP project and the SuperWASP cameras. *Publ. Astron. Soc. Pacif.* **118,** 1407–1418 (2006).
45. Bakos, G. *et al.* Wide-field millimagnitude photometry with the HAT: a tool for extrasolar planet detection. *Publ. Astron. Soc. Pacif.* **116,** 266–277 (2004).
46. Bodenheimer, P., Lin, D. N. C. & Mardling, R. A. On the tidal inflation of short-period extrasolar planets. *Astrophys. J.* **548,** 466–472 (2001).
47. Guillot, T. & Showman, A. P. Evolution of 51 Pegasus b-like planets. *Astron. Astrophys.* **385,** 156–165 (2002).
48. Burrows, A., Hubeny, I., Budaj, J. & Hubbard, W. B. Possible solutions to the radius anomalies of transiting giant planets. *Astrophys. J.* **661,** 502–514 (2007).
49. Chabrier, G. & Baraffe, I. Heat transport in giant (Exo)planets: a new perspective. *Astrophys. J.* **661,** L81–L84 (2007).
50. Batygin, K. & Stevenson, D. J. Inflating hot Jupiters with ohmic dissipation. *Astrophys. J.* **714,** L238–L243 (2010).
51. Triaud, A. H. M. J. *et al*. Spin-orbit angle measurements for six southern transiting planets. New insights into the dynamical origins of hot Jupiters. *Astron. Astrophys.* **524,** A25 (2010).
52. Winn, J. N., Fabrycky, D., Albrecht, S. & Johnson, J. A. Hot stars with hot Jupiters have high obliquities. *Astrophys. J.* **718,** L145–L149 (2010).
53. Rasio, F. A. & Ford, E. B. Dynamical instabilities and the formation of extrasolar planetary systems. *Science* **274,** 954–956 (1996).
54. Holman, M., Touma, J. & Tremaine, S. Chaotic variations in the eccentricity of the planet orbiting 16 Cygni B. *Nature* **386,** 254–256 (1997).
55. Crida, A. & Batygin, K. Spin-orbit angle distribution and the origin of (mis) aligned hot Jupiters. *Astron. Astrophys.* (in the press).
56. Santos, N. C. *et al*. The HARPS survey for southern extra-solar planets. II. A 14 Earth-masses exoplanet around μ Arae. *Astron. Astrophys.* **426,** L19–L23 (2004).
57. Butler, R. P. *et al*. A Neptune-mass planet orbiting the nearby M dwarf GJ 436. *Astrophys. J.* **617,** 580–588 (2004).
58. McArthur, B. E. *et al*. Detection of a Neptune-mass planet in the ρ-1 Cancri system using the Hobby-Eberly telescope. *Astrophys. J.* **614,** L81–L84 (2004).
59. Lovis, C. *et al*. An extrasolar planetary system with three Neptune-mass planets. *Nature* **441,** 305–309 (2006).
60. Lovis, C. *et al*. Towards the characterization of the hot Neptune/super-Earth population around nearby bright stars. *Proc. IAU Symp.* **253,** 502–505 (2009).
61. Borucki, W. J. *et al*. Characteristics of planetary candidates observed by Kepler II. Analysis of the first four months of data. *Astrophys. J.* **736,** 19 (2011).
62. Bonfils, X. *et al*. The HARPS search for southern extra-solar planets. XXXI. The M-dwarf sample. *Astron. Astrophys.* **549,** A109 (2013).
63. Howard, A. W. *et al*. The occurrence and mass distribution of close-in super-Earths, Neptunes, and Jupiters. *Science* **330,** 653–655 (2010).
64. Terquem, C. & Papaloizou, J. C. B. Migration and the formation of systems of hot super-Earths and Neptunes. *Astrophys. J.* **654,** 1110–1120 (2007).
65. Chiang, E. & Laughlin, G. The minimum-mass extrasolar nebula: *in situ* formation of close-in super-Earths. *Mon. Not. R. Astron. Soc.* **431,** 3444–3455 (2013).
66. Dawson, R. I. & Fabrycky, D. C. Radial velocity planets de-aliased: a new, short period for super-Earth 55 Cnc e. *Astrophys. J.* **722,** 937–953 (2010).
67. Demory, B.-O. *et al*. Detection of a transit of the super-Earth 55 Cancri e with warm Spitzer. *Astron. Astrophys.* **533,** A114 (2011).
68. Winn, J. N. *et al*. A super-Earth transiting a naked-eye star. *Astrophys. J.* **737,** L18 (2011).
69. Howard, A. W. *et al*. The NASA-UC Eta-Earth program. III. A super-Earth orbiting HD 97658 and a Neptune-mass planet orbiting Gl 785. *Astrophys. J.* **730,** 10 (2011).
70. Dragomir, D. *et al*. MOST detects transits of HD 97658b, a warm, likely volatile-rich super-Earth. *Astrophys. J.* **772,** L2 (2013).
71. Charbonneau, D. *et al*. A super-Earth transiting a nearby low-mass star. *Nature* **462,** 891–894 (2009).
72. Bonfils, X. *et al*. A hot Uranus transiting the nearby M dwarf GJ 3470. Detected with HARPS velocimetry. Captured in transit with TRAPPIST photometry. *Astron. Astrophys.* **546,** A27 (2012).
73. Gillon, M. *et al*. Detection of transits of the nearby hot Neptune GJ 436 b. *Astron. Astrophys.* **472,** L13–L16 (2007).
74. Bakos, G. A. *et al*. HAT-P-11b: A super-Neptune planet transiting a bright K star in the Kepler field. *Astrophys. J.* **710,** 1724–1745 (2010).
75. Marcy, G. W. *et al*. Masses, radii, and orbits of small Kepler planets: the transition from gaseous to rocky planets. *Astrophys. J.* **210,** 20 (2014).
76. Pepe, F. *et al*. The HARPS search for Earth-like planets in the habitable zone. I. Very low-mass planets around HD 20794, HD 85512, and HD 192310. *Astron. Astrophys.* **534,** A58–A73 (2011).
This paper reports the detection of several super-Earths with sub-metre per second Doppler signals, including one close to the habitable zone of a K dwarf.
77. Selsis, F. *et al*. Habitable planets around the star Gliese 581? *Astron. Astrophys.* **476,** 1373–1387 (2007).
78. Delfosse, X. *et al*. The HARPS search for southern extra-solar planets. XXXIII. Super-Earths around the M-dwarf neighbors Gl 433 and Gl 667C. *Astron. Astrophys.* **553,** A8 (2013).
79. Tuomi, M. *et al*. Habitable-zone super-Earth candidate in a six-planet system around the K2.5V star HD 40307. *Astron. Astrophys.* **549,** A48 (2013).
80. Kopparapu, R. K. *et al*. Habitable zones around main-sequence stars: new estimates. *Astrophys. J.* **765,** 131 (2013).
81. Dumusque, X., Santos, N. C., Udry, S., Lovis, C. & Bonfils X. Planetary detection limits taking into account stellar noise. II. Effect of stellar spot groups on radial-velocities. *Astron. Astrophys.* **527,** A82 (2011).
82. Dumusque, X. *et al*. The HARPS search for southern extra-solar planets. XXX. Planetary systems around stars with solar-like magnetic cycles and short-term activity variation. *Astron. Astrophys.* **535,** A55 (2011).
83. Meunier, N. & Lagrange, A.-M. Using the Sun to estimate Earth-like planets detection capabilities. IV. Correcting for the convective component. *Astron. Astrophys.* **551,** A101 (2013).
84. Aigrain, S., Pont, F. & Zucker, S. A simple method to estimate radial velocity variations due to stellar activity using photometry. *Mon. Not. R. Astron. Soc.* **419,** 3147–3158 (2012).
85. Boisse, I., Bonfils, X. & Santos, N. C. SOAP. A tool for the fast computation of photometry and radial velocity induced by stellar spots. *Astron. Astrophys.* **545,** A109 (2012).
86. Torres, G. *et al*. Improved parameters for extrasolar transiting planets. *Astrophys. J.* **677,** 1324–1342 (2008).
87. Guillot, T. *et al*. A correlation between the heavy element content of transiting extrasolar planets and the metallicity of their parent star. *Astron. Astrophys.* **453,** L21–L24 (2006).
88. Santos, N. C., Israelian, G. & Mayor, M. Spectroscopic [Fe/H] for 98 extra-solar planet-host stars. Exploring the probability of planet formation. *Astron. Astrophys.* **415,** 1153–1166 (2004).
A large-scale study of the correlation between giant-planet occurrence and the metallicity of the host star.
89. Fischer, D. A. & Valenti, J. The planet-metallicity correlation. *Astrophys. J.* **622,** 1102–1117 (2005).
90. Sousa, S. G. *et al*. Spectroscopic stellar parameters for 582 FGK stars in the HARPS volume-limited sample. Revising the metallicity-planet correlation. *Astron. Astrophys.* **533,** A141 (2011)

91. Buchhave, L. A. An abundance of small exoplanets around stars with a wide range of metallicities. *Nature* **486,** 375–377 (2012).
92. Adibekyan, V. Zh. *et al.* Overabundance of α-elements in exoplanet-hosting stars. *Astron. Astrophys.* **543,** A89 (2012)
93. Israelian, G. *et al.* Enhanced lithium depletion in Sun-like stars with orbiting planets. *Nature* **462,** 189–191 (2009).
94. Baumann, P. *et al.* Lithium depletion in solar-like stars: no planet connection. *Astron. Astrophys.* **519,** A87 (2010).
95. Israelian, G. *et al.* Evidence for planet engulfment by the star HD82943. *Nature* **411,** 163–166 (2001).
96. Reddy, B. *et al.* A search for ^{6}Li in stars with planets. *Mon. Not. R. Astron. Soc.* **335,** 1005–1016 (2002).
97. Ramírez, I. *et al.* A possible signature of terrestrial planet formation in the chemical composition of solar analogs. *Astron. Astrophys.* **521,** A33 (2010).
98. González Hernández, J. I. *et al.* Searching for the signatures of terrestrial planets in solar analogs. *Astrophys. J.* **720,** 1592–1602 (2010).
99. Dawson, R. & Murray-Clay, R. A. Giant planets orbiting metal-rich stars show signatures of planet-planet interactions. *Astrophys. J.* **767,** L24 (2013).
100. Adibekyan, V. Zh. *et al.* Orbital and physical properties of planets and their hosts: new insights on planet formation and evolution. *Astron. Astrophys.* **560,** A51 (2013).
101. Butler, R. P. *et al.* Evidence for multiple companions to υ Andromedae. *Astrophys. J.* **526,** 916–927 (1999).
102. Mazeh, T. *et al.* The spectroscopic orbit of the planetary companion transiting HD 209458. *Astrophys. J.* **532,** L55–L58 (2000).
103. Naef, D. *et al.* HD 80606 b, a planet on an extremely elongated orbit. *Astron. Astrophys.* **375,** L27–L30 (2001).
104. Moutou, C. *et al.* Photometric and spectroscopic detection of the primary transit of the 111-day-period planet HD 80 606 b. *Astron. Astrophys.* **498,** L5–L8 (2009).
105. Pepe, F. *et al.* The HARPS search for southern extra-solar planets. VIII. μ Arae, a system with four planets. *Astron. Astrophys.* **462,** 769–776 (2007).
106. Fischer, D. A. *et al.* Five planets orbiting 55 Cancri. *Astrophys. J.* **675,** 790–801 (2008).
107. Bouchy, F. *et al.* ELODIE metallicity-biased search for transiting hot Jupiters. II. A very hot Jupiter transiting the bright K star HD 189733. *Astron. Astrophys.* **444,** L15–L19 (2005).
108. Sato, B. *et al.* The N2K Consortium. II. A transiting hot Saturn around HD 149026 with a large dense core. *Astrophys. J.* **633,** 465–473 (2005).
109. Udry, S. *et al.* The HARPS search for southern extra-solar planets. XI. Super-Earths (5 and 8 M_{E}) in a 3-planet system. *Astron. Astrophys.* **469,** L43–L47 (2007).
110. Mayor, M. *et al.* The HARPS search for southern extra-solar planets. XVIII. An Earth-mass planet in the GJ 581 planetary system. *Astron. Astrophys.* **507,** 487–494 (2009).
111. Hébrard, G. *et al.* Misaligned spin-orbit in the XO-3 planetary system? *Astron. Astrophys.* **488,** 763–770 (2008).
112. Mayor, M. *et al.* The HARPS search for southern extra-solar planets. XIII. A planetary system with 3 super-Earths (4.2, 6.9, and 9.2 M_{E}). *Astron. Astrophys.* **493,** 639–644 (2009).
113. Rivera, E. J. *et al.* The Lick-Carnegie exoplanet survey: a Uranus-mass fourth planet for GJ 876 in an extrasolar Laplace configuration. *Astrophys. J.* **719,** 890–899 (2010).
114. Queloz, D. *et al.* WASP-8b: a retrograde transiting planet in a multiple system. *Astron. Astrophys.* **517,** L1 (2010).
115. Lovis, C. *et al.* The HARPS search for southern extra-solar planets. XXVIII. Up to seven planets orbiting HD 10180: probing the architecture of low-mass planetary systems. *Astron. Astrophys.* **528,** A112 (2011).
This article reports the discovery of a densely populated system with up to seven planets and the study of its dynamical architecture.
116. Triaud, A. H. M. J. *et al.* The Rossiter-McLaughlin effect of CoRoT-3b and HD 189733b. *Astron. Astrophys.* **506,** 377–384 (2009).
117. Zeng, L. & Sasselov, D. A detailed model grid for solid planets from 0.1 through 100 Earth masses. *Publ. Astron. Soc. Pacif.* **125,** 227–239 (2013).

Acknowledgements We thank A. Triaud for his help in preparing Fig. 2. N.C.S. was supported by Fundação para a Ciência e a Tecnologia (FCT, Portugal) through the Investigador FCT contract reference IF/00169/2012 and POPH/FSE (EC) by FEDER funding through the program Programa Operacional de Factores de Competitividade-COMPETE. N.C.S. further acknowledges the support from the European Research Council/European Community under FP7 through Starting Grant agreement number 239953. M.M and C.L. acknowledge the support of the Swiss National Science Foundation.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this paper at go.nature.com/z9q3xp. Correspondence should be addressed to M.M. (michel.mayor@unige.ch).

REVIEW

doi:10.1038/nature13781

Advances in exoplanet science from Kepler

Jack J. Lissauer[1], Rebekah I. Dawson[2] & Scott Tremaine[3]

Numerous telescopes and techniques have been used to find and study extrasolar planets, but none has been more successful than NASA's Kepler space telescope. Kepler has discovered most of the known exoplanets, the smallest planets to orbit normal stars and the planets most likely to be similar to Earth. Most importantly, Kepler has provided us with our first look at the typical characteristics of planets and planetary systems for planets with sizes as small as, and orbits as large as, those of Earth.

Kepler is a 0.95-m aperture space telescope launched by NASA in 2009 (refs 1, 2). Kepler identifies those exoplanets whose orbits happen to appear 'edge-on' by searching for periodic dips caused by planetary transits (partial eclipses) of the stellar disks. Above Earth's atmosphere, and in an Earth-trailing heliocentric orbit away from the glare and thermal variations of a low Earth orbit, Kepler monitored the brightness of more than 10^5 stars at 30-minute cadence for 4 years. Kepler's unique asset was its unprecedented photometric precision of around 30 parts per million for 12th magnitude stars with data binned in 6.5-hour intervals[3]. This is the benchmark time interval because Earth takes 13 hours to transit the Sun when viewed by a distant observer in the ecliptic plane (observers slightly away from the ecliptic view a transit of shorter duration). Such high-precision measurements are only possible in space (where stars do not twinkle), and are required in order to search for Earth-like planets — because the transit of such a planet across the disk of a Sun-like star blocks only 80 parts per million of the stellar flux. For comparison, the transit of a Jupiter-size planet across a similar star blocks 1% of the flux, and this dip is straightforward to detect using ground-based telescopes.

Transits of around 3,600 planet candidates, most of which represent true exoplanets, as described below, have been identified in the first 3 years of Kepler data (Fig. 1). The discovery of these planets, most of which have orbital periods (local 'years') shorter than a few Earth months, has greatly expanded the collection of known exoplanet types. Most Kepler planets have radii, R_p, that are intermediate between those of Earth and Neptune (1–3.8 R_E; where R_E = 6,371 km, Earth's radius); planets in this size range are missing from our Solar System. These planets have a wide range of densities[4–9], probably because they have atmospheres with a wide range of properties. Nonetheless, theoretical models of their interiors[10] imply that all of the planets in this class are 'gas-poor', that is, less than half — in most cases much less — of their mass consists of hydrogen and helium (H/He). By contrast, H/He make up more than 98% of our Sun's mass as well as substantial majorities of the masses of Jupiter, Saturn and almost all known giant exoplanets with $R_p > 9\,R_E$.

Kepler's primary mission is to conduct a statistical census of the abundance of planets as a function of planetary size, orbital period and stellar type. Kepler has found that planets are common, with the number of planets in the extended solar neighbourhood of our Galaxy being comparable with, or larger than, the number of stars[11]. Of particular interest is η_E, the average number of Earth-like planets per star. 'Earth-like' means having a radius similar to that of Earth and receiving about as much energy flux from its host star as Earth receives from our Sun; a more precise definition is given later. With some extrapolation downward in size and longward in orbital period, Kepler data suggest that η_E is ~0.1, although, as discussed later, there is a broad range of estimates for this value. Earlier studies[12,13] have found giant planets to be much more common around stars that are richer in heavy elements relative to light gases; Kepler data have shown that no comparable trend exists for small planets[14,15]. Almost half of Kepler's planet candidates are in systems in which multiple transiting planets have been found. As we discuss, the large abundance of such systems implies that flat systems containing multiple planets on closely spaced orbits are quite common. This finding supports models of planet formation within a disk of material orbiting a star first developed by Immanuel Kant and Pierre-Simon Laplace.

Because of Kepler's success, NASA extended the operations of the spacecraft beyond the original baseline plan, but Kepler's prime mission ended in May 2013 with the failure of a second reaction wheel that made precise stable pointing away from the spacecraft's orbital plane impossible. Nevertheless, data analysis over the next few years is expected to reveal hundreds or even thousands of planet candidates in addition to the several thousand already discovered, probably including some that extend the range of exoplanets to smaller sizes and longer periods (Fig. 1, bottom right), and perhaps including true Earth analogues, in terms of size and period, that orbit Sun-like stars. These additional planets, as well as better estimates of planetary sizes and planet detectability, will allow for improved estimates of the population of planets within our Galaxy. Although the partially disabled Kepler spacecraft cannot observe its original star field any longer, it has been reprogrammed to continue its search for other planets, with a focus on those orbiting small stars with orbital periods of less than 1 month; Kepler's new mission is dubbed 'K2'. Other space missions will continue to expand and exploit Kepler's discoveries over the next decade. The European Space Agency (ESA) launched the Gaia astrometric spacecraft in 2013, which will determine precise distances to stars hosting planets discovered by Kepler, enabling more accurate determination of the sizes of these stars and their associated planets. NASA's Transiting Exoplanet Survey Satellite (TESS), scheduled to launch in 2017, will conduct an all-sky search for transiting planets around the nearest and brightest stars using small-aperture, wide-field optics[16]. TESS planets will be easier to study with other space- and ground-based observatories than Kepler planets, most of which orbit much fainter stars. Searches for transiting planets in space will continue in the 2020s with ESA's Planetary Transits and Oscillations of Stars (PLATO)

[1]NASA Ames Research Center, Moffett Field, California 94035, USA. [2]Department of Astronomy, University of California, Berkeley, California 94720, USA. [3]Institute for Advanced Study, Princeton, New Jersey 08540, USA.

mission, which will have an effective aperture almost as large as Kepler does, together with a much larger field of view. Analysis of data from these advanced observatories, together with associated theoretical studies, should advance the studies of exoplanets that were pioneered by Kepler far beyond the mission's original goals.

Transiting planets and eclipsing binary stars

The transit depth yields the ratio of the planetary radius to the stellar radius, and the repetition rate of transits tells us the planet's orbital period. The stellar colours — or, better yet, stellar spectrum — can be used to deduce the star's radius and mass, and from these we can find the planet's radius and the semi-major axis of its orbit (from Kepler's third law). In favourable cases (generally restricted to close-in planets that are subject to intense stellar irradiation), we can detect the occultation of the planet as it travels behind the star, and thus determine the planet's albedo (its reflectivity). A wide range of albedos are found for both small[17] and large planets[18], with most hot giant planets having low albedo. Planets in multiple systems perturb one another through their mutual gravity, causing their orbits to deviate from strict periodicity. These deviations lead to transit timing variations (TTVs) that, in favourable cases, can be used to measure the planetary masses and additional orbital elements[5,7,8,19–21].

The objects in the catalogues assembled by the Kepler project[22–24] are considered to be only planet candidates because eclipsing binary stars can mimic transiting planets. Normally, the fractional brightness change in a binary-star eclipse is much larger than in a planetary transit, but occasionally the eclipse is grazing, or light from the Kepler-target star is diluted or 'blended' with the light from a background or companion eclipsing binary nearby on the plane of the sky. Such false-positives plague ground-based searches for exoplanets, but the Kepler light curves (starlight received as a function of time) are of such high quality that they can usually be used to discriminate between grazing or blended stellar eclipses and planetary transits. Moreover, Kepler is an imaging instrument, which can measure changes in the position of the image on the sky plane during transit. When multiple sources are present in the aperture used to measure the light curve from a Kepler target, the photometric centroid moves in response to flux changes in any of the sources. This property allows Kepler's pixel-level data to be used to test scenarios in which a planetary transit-like event is produced by a diluted background eclipsing binary star[25]. This 'centroiding' weeds out most eclipsing binaries that are blended with background stars and some that are blended with companion stars. Thus, although well under half of Kepler's planet candidates have been verified to be true planets, the false-positive rate of the catalogue as a whole is probably less than 10%; however, it may exceed 30% for the largest planet candidates[26–28]. Therefore, with appropriate care, the Kepler catalogue can be used for statistical studies of the exoplanet population.

Most of the planet candidates were found by the automated Kepler data-analysis pipeline, but many of the larger candidates were first identified by eye, including several dozen using data from the public Kepler archive by citizen scientists through the Planet Hunters project[29]. Other groups have found dozens of planets with orbital periods of less than 1 day, for which the pipeline is not optimized[30].

Planet candidates that have been verified to be true planets at a high level of confidence (in most cases well above 99%) are assigned Kepler designations (names). Verification can take the form of dynamical confirmation by detection of either TTVs in the Kepler light curve or radial-velocity variations, or it can be based on statistical arguments showing that the likelihood of the planet hypothesis is much greater than that of other possible causes of the observed light curve[31–33]. Typically, verified systems have been studied in greater detail than unverified candidates, resulting in more accurate estimates of stellar and planetary properties.

Individual planets and planetary systems

Kepler's primary mission was a statistical characterization of the exoplanet population. However, we first describe some of the highlights of the individual planets and planetary systems found.

Kepler's first major discovery was the Kepler-9 system, which contains

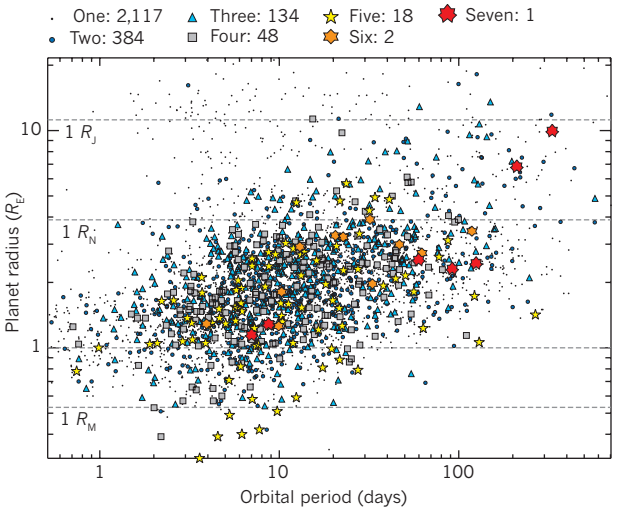


Figure 1 | Orbital period against planetary radius for planetary candidates. Coloured symbols represent the number of planetary candidates in the system found by analysing the first 3 years of Kepler data. The numbers shown above the figure represent the total number of systems of a given multiplicity in the catalogue; a small fraction of these planets fall outside the boundary of the period–radius ranges plotted. Planets with shorter orbital periods are over-represented because geometric factors and frequent transits make them easier for Kepler to detect. The upward slope in the lower envelope of these points is caused by the difficulty in detecting small planets with long orbital periods, for which transits are shallow and few are observed. The apparent absence of giant planets in multi-planet systems has been quantitatively confirmed[127]. R_E, radius of Earth; R_J, radius of Jupiter; R_M, radius of Mars; R_N, radius of Neptune. Data provided by J. Rowe.

two transiting giant planets with orbital periods (P) of 19.24 days and 38.91 days. The period ratio (38.91/19.24 = 2.02) is close to the 2:1 orbital resonance, which induces TTVs of tens of minutes in both planets. Analysis of these TTVs enabled both planets to be confirmed and provided estimates of their masses: each planet is similar in size to Saturn but less than half as massive[19,21]. TTVs have also been used to discover non-transiting planets, such as Kepler-19c[34] and Kepler-46c[35].

The first rocky planet found by Kepler[4] was Kepler-10b, which has an R_p of 1.42 ± 0.04 R_E and a mass, M_p, of 4.6 ± 1.2 M_E, where $M_E = 5.978 \times 10^{24}$ kg, Earth's mass. This planet's density, $8.8^{+2.1}_{-2.9}$ g cm^{-3}, is consistent with an Earth-like composition. Remarkably, its orbital period is only 20 hours.

Kepler-11 is a Sun-like star with six transiting planets that range in size from ~1.8–4.2 R_E[5,20]. Orbital periods of the inner five of these planets are between 10 and 47 days, with the ratio of orbital periods between adjacent planets ranging from 1.26 to 1.74. For comparison, the ratio of orbital periods in the Solar System ranges from 1.63 (Venus and Earth) to 6.3 (Mars and Jupiter). The outermost planet, Kepler-11g, has a period of 118.4 days. TTVs have been used to estimate the planets' masses. Most, if not all, have a substantial fraction of their volume occupied by the light gases hydrogen (H_2) and helium, implying that these gases can dominate the volume of a planet that is only a few times as massive as Earth. Kepler-36 hosts two planets, the semi-major axes of which differ by only 10%, but whose compositions are markedly different[7]: rocky Kepler-36b has an M_p of 4.5 ± 0.3 M_E, a density of $7.46^{+0.74}_{-0.59}$ g cm^{-3} and an orbital period of 13.84 days, whereas puffy Kepler-36c has an M_p of 8.7 ± 0.5 M_E, a density of $0.89^{+0.07}_{-0.05}$ g cm^{-3} and $P = 16.24$ days. It is possible that the atmosphere of Kepler-36b was stripped by photoevaporation or impact erosion, whereas Kepler-36c was able to retain its atmosphere because of its larger core mass and slightly larger distance from the host star[36]. The proximity of the orbits also presents a conundrum: although numerical integrations show that the current configuration may be long-lived, most nearby configurations are unstable on short timescales[37], so it is far from clear how these planets arrived at their current orbits.

The first transiting circumbinary planet to be discovered, Kepler-16b,

BOX 1

Kepler's pre-launch goals

Kepler's goals were to explore the structure and diversity of extrasolar planetary systems and thereby to:

- determine the frequency of Earth-size and larger planets in or near the habitable zone of a wide variety of spectral types of stars;
- determine the distributions of size and orbital semi-major axes of these planets;
- estimate the frequency of planets in multiple-star systems;
- determine the distributions of semi-major axes, albedo, size, mass and density of short-period giant planets;
- identify additional members of each photometrically discovered planetary system using complementary techniques;
- determine the properties of those stars that harbour planetary systems.

is an object of approximately Saturn's mass and radius ($M_p = 106 \pm 5\ M_E$, $R_p = 8.27 \pm 0.03\ R_E$), travelling on a nearly circular orbit (eccentricity $e = 0.0069$) with a period of 228.8 days around an eclipsing pair of stars with an orbital period of 41.08 days[6]. A bonus in such systems is that the planetary transits enable accurate measurements of the stellar masses and radii (errors $\lesssim$0.5%): one of the stars is about two-thirds the size and mass of our Sun and the other only a fifth as large as the Sun[6]. Moreover, the primary star's rotation axis has been measured to be aligned with the binary's orbital axis to within 2.4° (ref. 38). Several other circumbinary planets have been found using Kepler data, including the multi-planet Kepler-47 system[39].

Kepler-20e is the first planet smaller than Earth ($R_p = 0.87^{+0.08}_{-0.10}\ R_E$) to be verified around a main-sequence star other than the Sun[40]; its 6.1-day orbit means that it is much too hot to be habitable. The low-mass (M-dwarf) star Kepler-42 hosts three planets smaller than Earth, the smallest of which is Mars-sized[41]. Kepler-37b, only slightly larger than Earth's Moon, is the first planet smaller than Mercury to be found orbiting a main-sequence star; its period is 13 days, and the stellar host is 80% as massive as the Sun[42]. KIC 12557548b exhibits transits of varying depths, which might be due to an evaporating dusty atmosphere[43]. Kepler-78b[44] has the shortest orbital period of any confirmed exoplanet, circling its star in 8.5 hours. This roasting world is slightly larger than Earth, and its mass, measured from the radial-velocity variations it induces in its nearby host star, implies a rocky composition[45,46].

Circumstellar habitable zones are conventionally defined to be the distances from stars at which planets with an atmosphere similar to that of Earth receive the right amount of stellar radiation to maintain reservoirs of liquid water on their surfaces[47]. Kepler-62f is the first known exoplanet whose size ($1.41 \pm 0.07\ R_E$) and orbital position suggest that it could well be a rocky world with stable liquid water at its surface[48].

Principal goals of the Kepler mission

Discoveries like the ones mentioned have captured a great deal of attention in the scientific community and beyond. But Kepler is, in essence, a statistical mission, designed to discover large numbers of planets in a survey with well-characterized selection criteria. The stated goals of the Kepler mission before launch[49] are shown in Box 1.

In its 4-year prime mission, Kepler observed almost 200,000 stars, including about 140,000 dwarf or main-sequence stars that were monitored for a substantial majority of that time. In addition to its contribution to exoplanet science, Kepler has revolutionized the field of asteroseismology, which probes stellar interiors by observing the surface manifestations of oscillations that propagate within stars[50,51], and has dramatically advanced our understanding of eclipsing binary stars[52], as well as other areas of stellar physics. This Review only considers stellar properties indirectly through their contributions to assessing planetary characteristics.

How common are planets?

The Kepler catalogue of planets is uniquely valuable for studying the structure and properties of planetary systems: it is large enough that we can map out the distribution of planets in multiple parameters (such as orbital period, radius, multiplicity and properties of the host star); it has, at least in principle, well-defined selection criteria (in contrast with radial-velocity catalogues, which come from many different surveys and which usually do not include null results); and most of the parameter space that it explores, typical radii of 1–3 R_E and orbital periods of up to approximately 1 year (Fig. 1), is not easily accessible by other techniques.

One of the most fundamental statistics describing planetary systems is the probability distribution $f(R_p,P)d\ln R_p d\ln P$ that a member of a specified class of stars possesses a planet in the infinitesimal area element[53–56] $d\ln R_p d\ln P$. The integral of this distribution over a range in planetary radius and orbital period is the average number of planets per star (not to be confused with the fraction of stars with planets, which is smaller). Kepler determines this distribution for Sun-like stars with reasonable accuracy for $R_p \gtrsim 1\ R_E$ in the range $P \lesssim 50$ days, and for $R_p \gtrsim 2\ R_E$ in the range $P \lesssim 150$ days.

Occurrence rate calculations must carefully account for the completeness, reliability and threshold criteria of the Kepler catalogue[57,58], as well as random and systematic errors in host-star properties. For candidates with small transit depths or just a few transits, robust estimates of occurrence rates require calibration by injecting and recovering planetary signals in Kepler data[59–62]. The false-positive frequency distribution must be modelled simultaneously[28]. Revisions of host-star properties can dramatically alter the radius distribution of planets[11]. So far, no occurrence rate calculations contain all of these ingredients. Here we summarize key results from early analyses of the Kepler data set (see ref. 63 for a review).

The studies cited above find that the number of planets per unit log period is nearly flat for $R_p \lesssim 4\ R_E$ and $P > 10$ days, but rises by a factor of 2–5 between orbital periods of 10 days and 100 days for larger planetary radii. The number of planets drops sharply for orbital periods below 10 days for $R_p \lesssim 4\ R_E$ and below 2–3 days for giant planets. The occurrence rate of giant planets on small orbits is a factor of three lower than in radial-velocity surveys[28,55,64], perhaps because a significant fraction of giant planets are injected into small orbits through planet–planet gravitational interactions, and the relatively metal-poor Kepler stars host fewer and/or less massive planets, which are less likely to interact strongly[65]. At all periods the number per log radius grows as the radius shrinks, at least down to radii of 2 R_E. Below 2 R_E, the distribution per unit log radius plateaus at orbital periods out to 50 days and probably out to 100 days[56,62]. The average number of planets per star with $P < 50$ days is ~0.2 for $1\ R_E < R_p < 2\ R_E$ and ~0.4 for all radii $R_p > 1\ R_E$[56,61]. For stars cooler and less massive than the Sun, the average number of planets per star is even higher[11]: $0.49^{+0.07}_{-0.05}$ for $1\ R_E < R_p < 2\ R_E$ and $0.69^{+0.08}_{-0.06}$ for all radii $R_p > 1\ R_E$ with $P < 50$ days. The higher frequency is remarkable given that the fixed period cutoff at 50 days corresponds to a smaller semi-major axis in the less massive stars.

One difficulty in discussing η_E (the mean number of Earth-like planets per star) is that different authors use different definitions for 'Earth-like'. For solar-type stars the most natural definition is $\eta_E = f(1\ R_E, 1\ \mathrm{yr})$; for other stars we can replace $P = 1$ yr with the period corresponding to the same incident stellar flux. Roughly speaking, this is the number of planets per star in a range of a factor of e in radius and period centred on the Earth's radius and period. Unfortunately, determining η_E according to this definition requires an extrapolation downwards in size and longward in orbital period from the region where Kepler has a statistically reliable planet sample, which introduces considerable uncertainty. Applying this extrapolation to power-law fits[56] $f(R_p,P) \propto P^{\beta}$ of the distribution of planets in the 16-month Kepler catalogue[23] yields $\eta_E = 0.09$. An independent analysis of Kepler light curves[62] gives a consistent result, $\eta_E = 0.12 \pm 0.04$, after renormalizing by a factor of 2.1 to convert their definition to ours.

However, a follow-up study[66] using the same catalogue, but a more general form for the period and radius distribution, found a much smaller value $\eta_E = 0.02^{+0.02}_{-0.01}$. There are also other uncertainties: for example, none of the results for η_E discussed here model false positives (for planets of this size, the biggest contributor is larger planets orbiting faint stars that appear close to the Kepler target on the plane of the sky[28]). Moreover, the extrapolations involve a mix of planets with and without substantial volatile envelopes (see below), as well as a likely mix of formation histories, and therefore the extrapolation may not capture the true occurrence rate.

A related number for cool, low-mass stars is $0.155^{+0.138}_{-0.098}$ planets per star with $0.5\,R_E < R_p < 1.4\,R_E$ receiving 0.46–1.0 times the solar flux at Earth[11], which corresponds to a rate of $0.26^{+0.23}_{-0.16}$ in an interval equal to that which we use to define η_E.

The diverse physical properties of Kepler planets

Kepler has discovered more than 3,000 planet candidates with radii $R_p < 4\,R_E$. Planetary interior models show warm planets of this size to be 'gas-poor', defined here as composed of less than 50% H/He by mass. Transit surveys are well-suited to studying the physical properties of such planets because their radii are very sensitive to small amounts of gas in the atmosphere — for example, just 1% H/He added to a 1 M_E and 1 R_E solid core can inflate the planet to 2 R_E[67] — and moderately sensitive to the bulk composition of the core (for example, water compared with rock). Furthermore, the gas-poor planets found by Kepler are valuable because they sample a wide range of incident fluxes and therefore were presumably subject to a wide range of photoevaporation rates; many are found at short orbital periods, allowing mass measurements or meaningful upper limits through radial-velocity follow-up studies[9]; and some are found in compact systems with multiple transiting, low-density planets, whose short orbital periods and large radii allow sensitive mass measurements through TTVs.

Figure 2 shows the masses, radii and incident flux received by well-characterized planets less than 20 times as massive as Earth. The wide range in size of gas-poor planets of a given mass indicates a diversity of composition. It should be noted that most of the sub-Saturn exoplanets whose masses and radii are both known are Kepler discoveries.

Various processes can affect the composition of gas-poor planets, including coagulation from volatile-rich or volatile-poor planetesimals, accretion of gas from the protoplanetary nebula if the planet forms before its dispersal, outgassing of volatiles from the planet's interior, atmospheric escape (for example, by photoevaporation), and erosion or enrichment of the atmosphere and mantle through collisions with planetesimals. Distinguishing which of these processes are at play and their relative contributions is both a challenge and a motivation for interpreting the measurements from Kepler.

Several of the planets discussed earlier (highlighted in Fig. 2) have served as case studies to illuminate the properties of gas-poor planets. In particular, they sample a continuum of photoevaporation rates, which are a function of both incident stellar flux (as a proxy for the X-ray/ultraviolet radiation responsible for atmospheric erosion) and core mass. The middle four planets of Kepler-11, each of which contains ~4–15% H/He by mass, might represent the pristine, uneroded initial compositions of gas-poor planets, whereas the innermost planet in the system, Kepler-11b, is only 0.5% H/He by mass (or maybe devoid of light gases entirely if it is water-rich), perhaps because it has undergone considerable mass loss from its primordial atmosphere[20]. Kepler-10b, with an incident flux about 30 times that of Kepler-11b, may in turn have lost all of its atmosphere to photoevaporation. This speculation is based on the density of Kepler-10b, which can be matched by theoretical models that do not require a volatile component[4]. In addition to lifetime-integrated X-ray and ultraviolet flux, core mass is an important factor in determining the photoevaporation rate. A larger core mass for Kepler-36c may have enabled it to maintain its atmosphere against photoevaporation, which may have stripped its nearby neighbour Kepler-36b[36].

There is now a large collection of gas-poor Kepler planets with masses that are individually less precisely measured than the case studies above, but nonetheless statistically powerful when analysed as an ensemble. Dozens of masses have been measured through radial-velocity follow-up[9]. Using an approach that accounts for degeneracies between mass and eccentricity[68], more than 100 planetary masses were estimated from TTVs[69]. Furthermore, theoretical models imply that radii of warm planets in the size range 2–4 R_E depend much more on H/He percentage than on total planet mass[67]. Thus, the larger sample of thousands of Kepler candidates with $R_p < 4\,R_E$, even if lacking measured masses, informs us of planetary occurrence rates as a function of composition and their correlations with other properties such as the orbital period and the mass and chemical composition of the host star, which in turn constrain models for the formation and evolution of gas-poor planets.

Several radius ranges may indicate different regimes for planet formation and evolution (Fig. 3). For $R_p \lesssim 1.6\,R_E$, most of the small number of transiting planets that have measured masses are dense enough to be rocky; by contrast, larger planets seem to require a volatile component[70]. This radius range includes all planets with a period of less than 1 day[30]; these planets may once have had atmospheres that have now been stripped by impacts, photoevaporation, stellar winds and/or tidal forces. For $1.6R_E \lesssim R_p \lesssim 3R_E$, the mass–radius relationship is consistent with $M_p \propto R_p$, indicating that the typical planetary density decreases with increasing size[71,72]. This mass–radius relationship requires a substantial mass fraction of water or a small mass fraction (0.1–5%) in a H/He atmosphere[71,72]. The scatter in the mass–radius relationship exceeds the measurement errors[72], indicating some diversity in composition and/or atmospheric properties that possibly includes rare rocky planets without voluminous atmospheres. The presence of a H/He atmosphere substantially increases the temperature at the rocky surface; thus planets such as Kepler-22b, which has a radius of 2.4 R_E[73], are unlikely to be habitable. For $3R_E \lesssim R_p \lesssim 7R_E$, the planets are less dense than water,

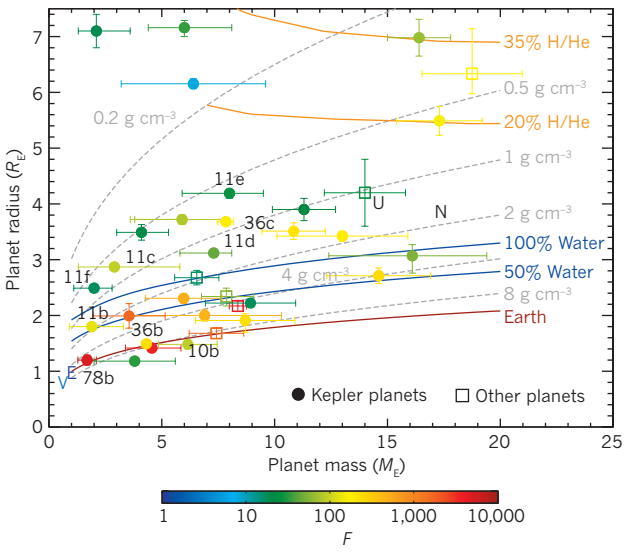


Figure 2 | Mass–radius plot for transiting exoplanets with measured masses less than 20 M_E, with the model curves for different compositions. Planets are colour-coded by the incident bolometric flux (F) they receive in units of the flux impinging on Earth. Kepler planets are shown by filled circles, with numbers and letters indicating planets discussed in this Review; the rocky planets in the crowded region near the lower left include Kepler-10b (red) and Kepler-36b (yellow). Other known exoplanets in this mass range are shown by open squares. The Solar System planets Venus (V), Earth (E), Uranus (U) and Neptune (N) are shown. The lower curve is for an Earth-like composition with two-thirds rock and one-third iron by mass. All other curves use thermal evolution calculations[128], assuming a volatile atmosphere of H/He or water atop a core of rock and iron with a composition the same as that of the bulk Earth. The two blue curves are for 50% and 100% water by mass and the two orange curves are for H/He atmospheres atop Earth-composition cores. These theoretical curves assume a radiation flux 100 times as large as that received by Earth and an age of 5 Gyr. Figure courtesy of E. Lopez.

implying voluminous H/He atmospheres[71]. The occurrence rate plummets between 2 R_E and 3 R_E[62,74], so this class is much rarer than the first two. Few planets in this class have been found around low-mass stars[71]. Large planets of $R_p \gtrsim 4\,R_E$ are more common around stars with larger abundances of elements heavier than helium[14,15].

The class of rocky planets with $R_p \lesssim 1.6\,R_E$ may lack gaseous atmospheres either because they were unable to accrete significant amounts of light gas and never outgassed an atmosphere, or because their primordial atmospheres were removed by impacts or photoevaporation. A possible explanation for the plunge in occurrence rate between 2 R_E and 3 R_E is that larger planets need H/He envelopes, which are uncommon, but most low-density smaller planets contain substantial water components or very low-mass outgassed H_2 envelopes. The paucity of planets larger than 3 R_E around low-mass stars and the higher heavy-element abundance in the host stars of systems containing planets larger than 4 R_E may both reflect the difficulty of accreting H/He envelopes in a protoplanetary disk with a low surface density in solids.

Properties of planetary systems

The Kepler catalogue is particularly important because it contains many multiple-planet systems, roughly eight times as many as all radial-velocity surveys combined[24]. Multiple systems are expected to have a very low false-positive rate ($\lesssim$1%), because background binary-star eclipses may mimic the light curve from a single transiting planet, but are unlikely to imitate two or more[33,75]. Multiple systems are also valuable because gravitational interactions among the planets lead to TTVs that in some cases allow us to determine the masses and orbital properties of one or more of the planets[5,7,8,19–21].

Multiple systems also allow us to constrain the average mutual inclinations of the planetary orbits, either by comparing relative transit durations and orbital periods[76] or by comparing the frequencies of systems with different multiplicities in the Kepler survey to those in radial-velocity surveys (since the chance that multiple planets in a single system will transit is much higher if their mutual inclination is low[77]). Such studies show that the typical mutual inclinations in Kepler planets are only a few degrees, similar to those in the Solar System. Another probe is the mutual inclination between a transiting planet and its non-transiting perturber, which is not biased towards low mutual inclinations by the selection effects that are present in multi-transiting systems; the first such systems with good constraints have been found to be flat[35,78].

The finding that typical Kepler multi-planet systems are flat is perhaps the first direct evidence that most planetary systems formed from a rotating thin disk of gas and dust, as suggested by Pierre-Simon Laplace more than two centuries ago. But even this widely accepted result leads to tension with other observations. In most formation models, planets have mean inclinations that are at least half as large as the mean eccentricities[79], and this result also holds for the planets in the Solar System, the asteroids and the Kuiper belt. Thus, we expect the mean eccentricity of the Kepler planets to be no more than about 0.1. Unfortunately, attempts to measure the eccentricity distribution of Kepler planets have been complicated by (and sometimes brought to light) systematic uncertainties in the stellar properties[80–82], although individual constraints have been possible for a subset of well-characterized stars with high signal-to-noise transits[83,84]. By contrast, the eccentricities of radial-velocity planets are straightforward to measure, and the mean eccentricity for those having orbital periods larger than 10 days is 0.26, much larger than we would expect from the arguments above. Are the eccentricities and inclinations of the radial-velocity planets larger than those of the Kepler planets? Or perhaps just larger than those of planets in Kepler's multiple planet systems? And if so, why? Could the eccentricities be over-estimated[85]? Or could exoplanets have much larger eccentricities than inclinations[86]?

An equally serious tension is revealed by ground-based measurements of the stellar obliquity, the angle between the equator of the host star and the orbital plane of a transiting planet. About 80 obliquities — or at least their projections on the sky plane — have been determined, mostly through measurements of the Rossiter–McLaughlin effect[87]. Almost half of the measured projected obliquities exceed 20°, and 15% exceed 90°; of course the width of the distribution of true (as opposed to projected) obliquities must be even larger. This result is quite different from the expectation for a Laplace-type model, in which the host star and planets form from a single rotating gas disk, and thus should have a common spin and orbital axis. One possibility is that the close-in giant planets arrived on their present orbits through high-eccentricity migration, which excites large obliquities[88]. Another possibility is that the stellar spin is misaligned with the axis of the planetary disk, perhaps because of a collision with a giant planet on a highly eccentric orbit or twisting of the planetary disk by external torques[89–91]. To complicate the situation further, most of the handful of Kepler planets for which measurements are available, including multi-planet systems, have obliquities near zero[50,92,93].

The properties of multi-planet systems are constrained by the requirement that they be dynamically stable over timescales comparable with the lifetime of the star. Rigorous stability criteria are not usually available except for two-planet systems[94], but a useful approximate criterion is that systems composed of planets on nearly circular, nearly co-planar orbits are stable for N orbits if the separation in semi-major axis between adjacent planets, $a_{i+1} - a_i$, exceeds some constant K_N times the mutual Hill radius (the separation at which the mutual planetary gravity equals the difference in the pull of the star on the two planets),

$$R_{H,i,i+1} \equiv \left(\frac{M_i + M_{i+1}}{3M_\star}\right)^{1/3} \frac{a_i + a_{i+1}}{2}$$

where M_i and M_{i+1} are the planet masses and $M_\star$ is the mass of the host star. For $N = 10^{10}$, $K_N \approx 9$–12 (ref. 95). As one would expect, most of the Kepler multi-planet systems are safely stable by this criterion (Fig. 4), and numerical integrations assuming initially circular, co-planar orbits confirm that almost all of them are stable for at least 10^{10} orbits[76,96]. These results depend on the assumed mass–radius relationship, but are relatively insensitive to it because the planet mass enters the definition of the Hill radius to the 1/3 power.

A deeper question is whether these systems are dynamically 'full' or 'packed', which we define to mean that no additional planets, even with very small masses, could be inserted between the existing ones in a stable orbital configuration. The situation in our own Solar System is ambiguous: the region from Jupiter to Neptune is packed, or nearly so[97], but inside of Mercury and between Earth and Mars there are significant bands in semi-major axis where additional low-mass planets would be stable for at least 10^8 years[98].

Dynamically packed systems are a natural consequence of models in which planets grow hierarchically, since planet growth should stop once all of the orbits are stable. However, the correspondence is not exact, since stable zones may not be occupied if the planets they contained collided in the final stages of hierarchical growth; moreover, the system may contain additional planets that are not transiting or fall below Kepler's detection threshold. The evidence (Fig. 4) suggests that the known Kepler planets are typically separated by about twice the distance required for stability, but given the possible presence of undiscovered planets and the destabilizing effects of non-zero eccentricity, many of these systems may be dynamically packed[99].

Planets are also found in binary star systems, either orbiting one of the two stars with an orbital period much shorter than the binary period ('S-type') or orbiting both with a period much longer than the binary period ('P-type' or 'circumbinary'). Most of our understanding of S-type planets comes from radial-velocity studies, whereas circumbinary planets around normal stars were first discovered by Kepler. In most respects the properties of planetary systems in single and binary-star systems are similar[100,101], although S-type planets seem to be less common in binary systems than similar planets around single stars[102].

Binary systems offer unique insights into the formation of both planets and stars. Torques from the companion star in an S-type binary can excite

slow, large-amplitude Lidov–Kozai oscillations in the eccentricity and inclination of the planetary orbit. One striking hint that Lidov–Kozai oscillations are sometimes at work is that the four planets with the largest eccentricities ($e > 0.85$) are all members of wide S-type binary systems[103]. A close companion star truncates the protoplanetary disk and the planetary system at a radius of about 0.25–0.3 times the companion's separation (for equal-mass stars on a circular orbit[104]). No S-type planets have been discovered, either in radial-velocity or transit surveys, in binary systems with separation less than about 10 times the Earth–Sun distance. This could mean that the site of planet formation in the protoplanetary disk is beyond 3 times (0.3×10) the Earth–Sun distance, consistent with theories in which planets found at smaller radii have migrated inward; alternatively, the outermost regions at which circumstellar orbits are stable may nonetheless be too perturbed for planets to form. If binary stars form through dynamical interactions between single stars in a dense gas-free cluster, it would be difficult for them to acquire circumbinary planets. However, if binaries form through fragmentation and collapse in a gas-rich environment, they are likely to acquire a circumbinary disk in which planets could form. Binary stars with orbital periods of a few days are likely to be formed from binaries with much longer periods through high-eccentricity migration induced by a tertiary companion[88]. This process would probably remove or destroy any planets initially orbiting one of the two stars and is unlikely to produce circumbinary planets detectable by Kepler. Therefore, we should not expect to find planets, S-type or P-type, in binary systems with periods of a few days or less, and this expectation is so far confirmed by the observations — the shortest-period planet-hosting binary star is Kepler-47 with a period of 7.45 days[39].

Just as important as the discoveries made by Kepler are its non-discoveries. So far, Kepler has found no co-orbital planets, which share the same average semi-major axis — like the Trojan asteroids found accompanying Jupiter and the Saturnian satellites Janus and Epimetheus. It has also found neither exomoons nor 'binary' planets orbiting one another[105,106].

Planet formation

The mass fraction of stellar material other than the dominant elements of hydrogen and helium (the metallicity, in astronomical parlance) ranges from a few per cent down to around 0.01% among stars in the solar neighbourhood. The initial protostellar disk presumably has the same composition as its host star, and these heavier elements are the raw material from which most of the mass in a typical Kepler planet must be drawn. Thus it is natural to expect that planet formation should be easier around stars having high metallicity. This expectation is confirmed for the giant planets detected in radial-velocity surveys — the fraction of high-metallicity stars hosting giant planets is much larger than the fraction of low-metallicity stars[12,13,107,108]. Similarly, a star in the Kepler catalogue with super-solar metallicity is around 2.5 times more likely to host a large planet ($R_p > 5\ R_E$) than a star with sub-solar metallicity[15]. Remarkably, there is no such correlation for small planets ($R_p < 2\ R_E$). The probabilities that a Kepler star with super-solar or sub-solar metallicity hosts a small planet are approximately equal[15], with a significant number of small Kepler planets found around stars with metallicity as small as one-quarter that of the Sun[14]. Perhaps this is a hint that the formation process for small planets has more than enough metals to draw on even in moderately low-metallicity disks. Before any firm conclusions are drawn we need reliable metallicities for a larger sample of Kepler stars and planet-frequency measurements for stars with a wider range of metallicities.

One of the most basic questions about the Kepler planets is whether they formed *in situ*[109,110] or whether they migrated to their current orbits from larger radii[111,112]. The orbital periods of most of Kepler's planets are $\lesssim 50$ days, resulting in nominal formation timescales that are much shorter than the lifetime of the gas disk, unlike those of the Solar System's terrestrial planets. Thus the planets and their gaseous envelopes could have formed *in situ*. The main argument for migration is that it is a robust process[113] that inevitably occurs in both analytical models and numerical simulations of planets orbiting in gaseous disks[114]. However, models of migration have not successfully predicted any populations of planets before they were observed.

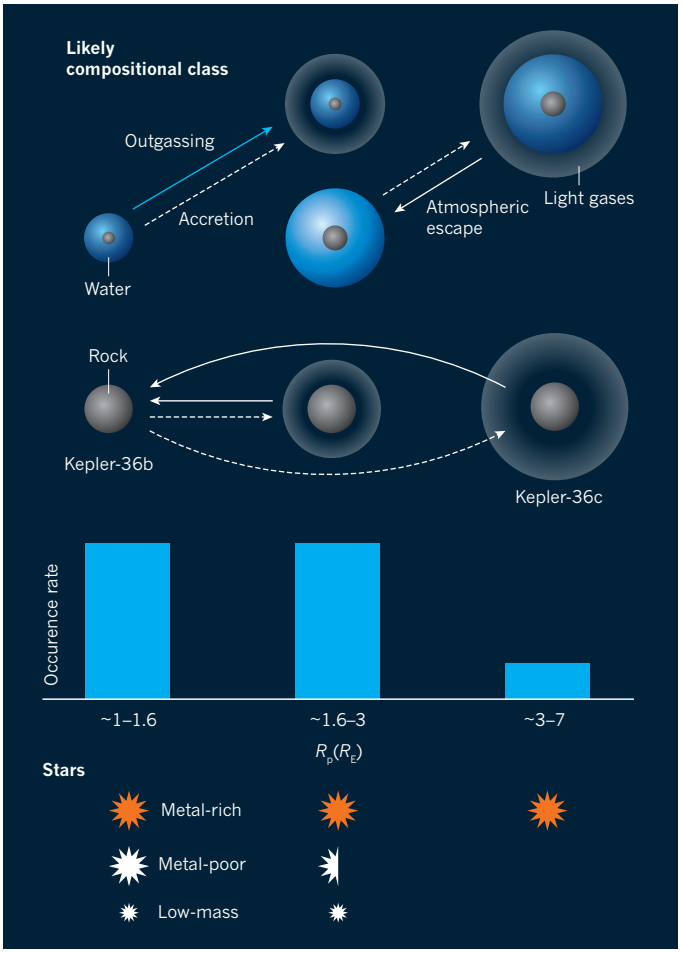


Figure 3 | Plausible compositions of the small and mid-sized planets observed by Kepler. The bars indicate the approximate relative occurrence rates[28,56,62] of planets in the specified size range. Plausible compositions in each size range, in terms of rock, water and light gases, are illustrated above. The arrows indicate physical processes that set or alter planet compositions. The smallest planets, on the left, can be rocky or possibly mixtures of rock and water. Somewhat larger planets have volumetrically significant amounts of constituents less dense than rock. Planets whose sizes are comparable to or larger than that of Neptune, $R_p = 3.8\ R_E$, have envelopes composed of the lightest gases, H_2 and He. The stars at the bottom of the figure indicate the masses and compositions of the stellar types that commonly host planets of the sizes indicated, with the half star indicating that although metal-poor stars host a significant number of planets in this size range, they are less common hosts.

Additional insight into the migration process comes from planets in mean-motion resonances. In the strongest of these, the orbital periods of the two resonant planets are in the ratio $(n + 1):n$, where n is an integer. Planets can be permanently captured into resonance if they cross the resonance during migration and the migration is convergent, that is, in a direction such that the period ratio evolves towards unity, rather than away. Capture into resonance during convergent migration is certain if the migration is slow enough and the planetary eccentricities are small enough[115]. It is therefore striking that the multi-planet systems discovered by Kepler contain very few resonant planet pairs. The excess fraction of planet pairs in the Kepler sample having period ratios within 5–10% of 3:2 or 2:1 is less than 5%. Possible explanations for the small fraction of resonant planets include the following: migration was too fast for capture to occur (this, however, requires migration times of $\lesssim 10^3$ years, much shorter than is plausible with disk migration[116]); stochastic torques on the migrating planet, which might arise in a turbulent protoplanetary disk[116], allowed the planets to escape the resonances and continue migrating; eccentricity damping from the protoplanetary

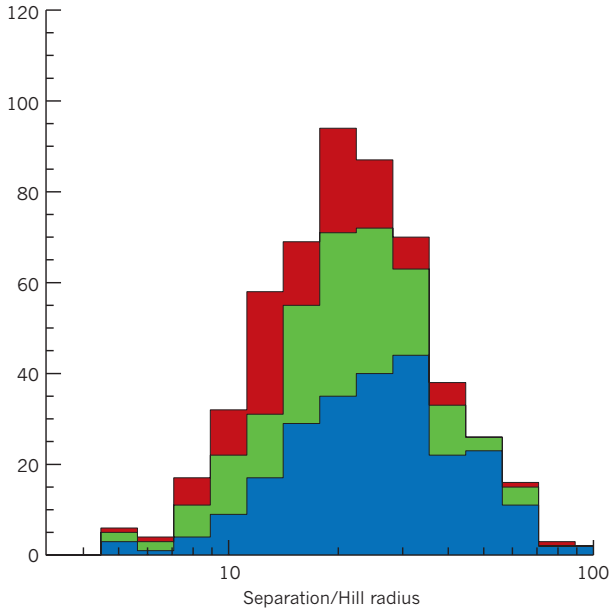


Figure 4 | Separations of nearest neighbours in the Kepler multi-planet systems, measured in Hill radii. Masses are derived using the mass–radius relationship $M_p = M_E (R_p/R_E)^{\alpha}$, where $\alpha = 3$ for $R_p < R_E$ and $\alpha = 2.06$ for $R_p > R_E$. Blue histogram, planet pairs in systems with two known planets; green histogram, planet pairs in systems with three known planets; red histogram, planet pairs in systems with four or more known planets. Planets on circular, co-planar orbits separated by more than 9–12 Hill radii are expected to be stable for the lifetime of typical stars; the few planet candidates seen with smaller separations may have large errors in their estimated radii, not obey the assumed mass–radius relationship or not be planets orbiting the same star. (Data from the catalogue of ref. 76.)

disk led to escape from the resonance[117] (this mechanism requires that the Kepler planets have very small eccentricities because eccentricities are difficult to excite after migration is complete[118]); and the planets might have formed *in situ* rather than migrating. Although still poorly understood, a possible clue to the answer comes from the distribution of period ratios in the Kepler multi-planet systems; these are asymmetric around the strong 2:1 and 3:2 resonances, with a peak just outside the resonance and/or a valley inside[76,119,120].

Unsolved problems

Kepler represents a watershed in our understanding of exoplanets and a great stride forward in understanding the properties of planetary systems and the problems in developing theories of their formation. But there are many aspects of planetary systems that Kepler has not illuminated. Kepler has opened up a new region in the orbital period versus radius plane (Fig. 1), containing planets as small as Earth's Moon at short periods, and larger planets with orbital periods as large as 1–2 years, but all of the planets in the Solar System lie outside this region (although only just outside in the case of Venus and Earth). The atmospheres of giant planets must be investigated through transit observations by ground-based telescopes or the Hubble and Spitzer space telescopes, as Kepler has no spectral resolution. The eccentricities of planetary orbits provide important insights into their formation, but only ground-based radial-velocity surveys can routinely measure eccentricities. However, these are expensive because the Kepler host stars are so faint, and often impossible with current technology because of the small masses of typical Kepler planets. Kepler cannot detect multiple transits of planets with periods longer than a few years, so Kepler has not advanced our understanding of the region beyond a few times the Earth–Sun distance, where most giant planets are likely to be born. Planets are occasionally found at much larger radii and many have probably been ejected into interstellar space, but these regions can only be investigated by high-resolution imaging and gravitational microlensing.

Kepler has hugely advanced our understanding of the phenomenology of exoplanets, but so far has led us no closer to a secure theory of planet formation. Did the Kepler planets form *in situ* or did they migrate from larger radii? Why are small planets common around host stars with such a wide range of metallicities? How did the Kepler planets acquire their voluminous atmospheres, and why are the atmospheres so diverse in mass even for planets with similar core masses? How are the Kepler planets related to the terrestrial planets in the Solar System? Why does the typical inclination of the Kepler planets, most of which are small, seem to be much less than the typical eccentricity of the radial-velocity planets, most of which are large? How are the large angles between some planetary orbital planes and the host-star equators generated? What is the relationship between the dynamics and formation of small, rocky planets and gas-giant planets[121]?

After Kepler

Although data acquisition by the Kepler spacecraft on its original target field has ended, ongoing data analysis and observational follow-up will refine the results already obtained and address some of the outstanding questions reviewed here.

Better models of the stellar and instrumental noise in the transit light curves, including a rigorous treatment of temporally correlated noise, might enable the discovery of smaller, longer-period planets[122,123] as well as better characterization of existing ones. More accurate transit times are crucial given that most Kepler planet masses are derived from TTVs, and more secure detections of, and upper limits on, transit duration variations (TDVs) will provide important constraints on mutual inclinations.

We can hope to address some of the unanswered questions about occurrence rates and system architectures by more sophisticated statistical analyses. The most complete view of the architectures of Kepler planetary systems will require tying together constraints from occurrence rates, transit durations, TTVs and TDVs. Even the absence of TTVs can provide important constraints on planetary system architectures — for example, close-in giant planets seem to have fewer (or no) companion planets compared with more distant or smaller planets[124]. Insight can be gained into the composition of gas-poor planets by joint modelling in the space of radius, incident flux and, where available, mass (Fig. 2).

Observational follow-up of Kepler targets from the ground and space is underway. High-quality spectra of Kepler host stars will help to refine estimates of their properties, and therefore reduce the uncertainty in stellar properties, which often dominates the uncertainty in planetary properties. A spectrum can also determine the star's projected rotational velocity which, combined with the stellar spin period from Kepler photometry, constrains the angle between the stellar equator and planetary orbit[93]. Spectra and adaptive-optics imaging will allow the catalogues to be culled more completely of false positives. The determination of accurate host-star metallicities, which has already provided new insights into the planet-formation process[14,15], will be greatly expanded by the Large Sky Area Multi-Object Fibre Spectroscopic Telescope (LAMOST)[125]. Programmes are being developed to follow up some Kepler candidates with large TTVs and establish a longer baseline for TDVs using ground-based telescopes. Kepler stars are among the billions astrometrically monitored by the Gaia mission. Gaia will determine the distances of the Kepler target stars, thereby improving our knowledge of stellar parameters and, consequently, the planetary radii; better radii will, in turn, improve estimates of the planetary occurrence rates and compositions, and correlations between planetary and stellar properties. For stars within around 200 pc, both within and outside the Kepler field, Gaia can detect Jupiter analogues by the astrometric oscillations of their host stars, revealing a more complete architecture for systems in which only the close-in planets are detectable by transits or radial-velocity measurements.

Space-based all-sky surveys, including TESS and PLATO (scheduled to launch by 2024) could greatly increase the science returns from Kepler by revisiting the Kepler field. Such follow-up would provide a long time baseline that would allow for improved occurrence rates, including an accurate value for η_E, masses for longer period planets from TTVs that would help to address outstanding issues about their formation and composition, and

the possibility of a substantial number of TDV detections.

Two spacecraft scheduled to be launched this decade will study known transiting planets. ESA's Characterising Exoplanets Satellite (CHEOPS) will target known exoplanet hosts to discover additional transiting gas-poor planets. NASA's James Webb Space Telescope (JWST) will characterize the atmospheres of gas-poor planets analogous to those that Kepler discovered in abundance; this might break the degeneracy among several composition possibilities.

Since the first handful of exoplanet discoveries two decades ago, the pace of exoplanet research has been extraordinary[126], driven primarily by ground-based radial-velocity searches of growing power and sophistication and by the Kepler mission. The armada of projects described in this Review will probe new regions of exoplanet parameter space and provide more detailed and accurate probes of the properties of known exoplanets. The challenges for the next two decades will be to maintain the momentum that has been built up in exoplanet research, and to work towards the even longer-term goal of producing an image of an extrasolar planet comparable with the iconic images of Earth taken by the Apollo astronauts. ■

Received 30 June; accepted 24 July 2014.

1. Borucki, W. J. *et al*. Kepler planet-detection mission: introduction and first results. *Science* **327,** 977–980 (2010).
   This is the primary paper describing the Kepler mission, its goals and first planetary discoveries.
2. Koch, D. G. *et al*. Kepler mission design, realized photometric performance, and early science. *Astrophys. J.* **713,** L79–L86 (2010).
   This article provides a description of the spacecraft, how the hardware relates to the scientific goals of the mission, and early technical performance.
3. Gilliland, R. L. *et al*. Kepler mission stellar and instrument noise properties. *Astrophys. J. Suppl*. **197,** 6 (2011).
4. Batalha, N. M. *et al*. Kepler's first rocky planet: Kepler-10b. *Astrophys. J.* **729,** 27 (2011).
5. Lissauer, J. J. *et al*. A closely packed system of low-mass, low-density planets transiting Kepler-11. *Nature* **470,** 53–58 (2011).
   This paper reports the first flat, tightly packed, close-in planetary system and the first small planets found with low densities.
6. Doyle, L. R. *et al*. Kepler-16: a transiting circumbinary planet. *Science* **333,** 1602–1606 (2011).
   This article describes the first transiting circumbinary planet.
7. Carter, J. A. *et al*. Kepler-36: A pair of planets with neighboring orbits and dissimilar densities. *Science* **337,** 556–559 (2012).
   The authors of this article describe a system of two planets with very different densities on remarkably close orbits that is precisely characterized using TTVs.
8. Jontof-Hutter, D., Lissauer, J. J., Rowe, J. F. & Fabrycky, D. C. Kepler-79's low density planets. *Astrophys. J.* **785,** 15 (2014).
9. Marcy, G. W. *et al*. Masses, radii, and orbits of small Kepler planets: the transition from gaseous to rocky planets. *Astrophys. J. Suppl*. **210,** 20 (2014).
   This paper describes the masses of dozens of small Kepler planets using radial velocities measured at the Keck Observatory.
10. Fortney, J. J., Marley, M. S. & Barnes, J. W. Planetary radii across five orders of magnitude in mass and stellar insolation: application to transits. *Astrophys. J.* **659,** 1661–1672 (2007).
11. Dressing, C. D. & Charbonneau, D. The occurrence rate of small planets around small stars. *Astrophys. J.* **767,** 95 (2013).
12. Fischer, D. A. & Valenti, J. The planet-metallicity correlation. *Astrophys. J.* **622,** 1102–1117 (2005).
13. Sousa, S. G., Santos, N. C., Israelian, G., Mayor, M. & Udry, S. Spectroscopic stellar parameters for 582 FGK stars in the HARPS volume-limited sample. Revising the metallicity-planet correlation. *Astron. Astrophys.* **533,** A141 (2011).
14. Buchhave, L. A. *et al*. An abundance of small exoplanets around stars with a wide range of metallicities. *Nature* **486,** 375–377 (2012).
   This article demonstrates that the occurrence rate of small planets does not depend strongly on the chemical composition of the star.
15. Wang, J. & Fischer, D. A. The metal-rich stars get richer in planets for all but planets with $R_P \leq 2\ R_E$. Preprint at http://arxiv.org/abs/1310.7830 (2013).
16. Ricker, G. R. *et al*. The Transiting Exoplanet Survey Satellite mission. *J. Am. Assoc. Variable Star Observers* **42,** 234 (2014).
17. Demory, B.-O. The albedos of Kepler's close-in super-Earths. *Astrophys. J.* **789,** L20 (2014).
18. Angerhausen, D., DeLarme, E. & Morse, J. A. A comprehensive study of Kepler phase curves and secondary eclipses — temperatures and albedos of confirmed Kepler giant planets. Preprint at http://arxiv.org/abs/1404.4348 (2014).
19. Holman, M. J. *et al*. Kepler-9: a system of multiple planets transiting a Sun-like star, confirmed by timing variations. *Science* **330,** 51–54 (2010).
20. Lissauer, J. J. *et al*. All six planets known to orbit Kepler-11 have low densities. *Astrophys. J.* **770,** 131 (2013).
21. Dreizler, S. & Ofir, A. Kepler-9 revisited 60% the mass with six times more data. http://arxiv.org/abs/1403.1372 (2014).
22. Borucki, W. J. *et al*. Characteristics of planetary candidates observed by Kepler. II. Analysis of the first four months of data. *Astrophys. J.* **736,** 19 (2011).
   This is the first major catalogue of Kepler's planet candidates.
23. Batalha, N. M. *et al*. Planetary candidates observed by Kepler. III. Analysis of the first 16 months of data. *Astrophys. J. Suppl*. **204,** 24 (2013).
   This is the second major catalogue of Kepler's planet candidates.
24. Burke, C. J. *et al*. Planetary candidates observed by Kepler IV: planet sample from Q1–Q8 (22 months). *Astrophys. J. Suppl*. **210,** 19 (2014).
   This is the third major catalogue of Kepler's planet candidates.
25. Batalha, N. M. *et al*. Pre-spectroscopic false-positive elimination of Kepler planet candidates. *Astrophys. J.* **713,** L103–L108 (2010).
26. Morton, T. D. & Johnson, J. A. On the low false positive probabilities of Kepler planet candidates. *Astrophys. J.* **738,** 170 (2011).
27. Santerne, A. *et al*. SOPHIE velocimetry of Kepler transit candidates. VII. A false-positive rate of 35% for Kepler close-in giant candidates. *Astron. Astrophys*. **545,** A76 (2012).
28. Fressin, F. *et al*. The false positive rate of Kepler and the occurrence of planets. *Astrophys. J.* **766,** 81 (2013).
   This paper estimates false-positive rates in Kepler catalogues and calculates planetary occurrence rates.
29. Fischer, D. A. *et al*. Planet hunters: the first two planet candidates identified by the public using the Kepler public archive data. *Mon. Not. R. Astron. Soc*. **419,** 2900–2911 (2012).
30. Sanchis-Ojeda, R. *et al.* A study of the shortest-period planets found with Kepler. *Astrophys. J*. **787,** 47 (2014).
   This is a catalogue of Kepler planet candidates with an orbital period <1 day.
31. Torres, G. *et al*. Modeling Kepler transit light curves as false positives: rejection of blend scenarios for Kepler-9, and validation of Kepler-9d, a super-Earth-size planet in a multiple system. *Astrophys. J.* **727,** 24 (2011).
32. Morton, T. D. An efficient automated validation procedure for exoplanet transit candidates. *Astrophys. J*. **761,** 6 (2012).
33. Lissauer, J. J. *et al.* Validation of Kepler's multiple planet candidates. II. Refined statistical framework and descriptions of systems of special interest. *Astrophys. J.* **784,** 44 (2014).
34. Ballard, S. *et al*. The Kepler-19 system: a transiting 2.2 $R_\oplus$ planet and a second planet detected via transit timing variations. *Astrophys. J*. **743,** 200 (2011).
35. Nesvorny, D. *et al*. The detection and characterization of a nontransiting planet by transit timing variations. *Science* **336,** 1133–1136 (2012).
36. Lopez, E. D. & Fortney, J. J. The role of core mass in controlling evaporation: the Kepler radius distribution and the Kepler-36 density dichotomy. *Astrophys. J*. **776,** 2 (2013).
37. Deck, K. M. *et al*. Rapid dynamical chaos in an exoplanetary system. *Astrophys. J. Lett*. **755,** L21 (2012); erratum **774,** L15 (2013).
38. Winn, J. N. *et al.* Spin-orbit alignment for the circumbinary planet host Kepler-16 A. *Astrophys. J*. **741,** L1 (2011).
39. Orosz, J. A. *et al*. Kepler-47: a transiting circumbinary multiplanet system. *Science* **337,** 1511–1514 (2012).
40. Fressin, F. *et al*. Two Earth-sized planets orbiting Kepler-20. *Nature* **482,** 195–198 (2012).
41. Muirhead, P. S. *et al*. Characterizing the cool KOIs. III. KOI 961: a small star with large proper motion and three small planets. *Astrophys. J*. **747,** 144 (2012).
42. Barclay, T. *et al.* A sub-mercury-sized exoplanet. *Nature* **494,** 452–454 (2013).
43. Rappaport, S. *et al*. Possible disintegrating short-period super-mercury orbiting KIC 12557548. *Astrophys. J*. **752,** 1 (2012).
44. Sanchis-Ojeda, R. *et al*. Transits and occultations of an Earth-sized planet in an 8.5 hr orbit. *Astrophys. J.* **774,** 54 (2013).
45. Howard, A. W. *et al.* A rocky composition for an Earth-sized exoplanet. *Nature* **503,** 381–384 (2013).
46. Pepe, F. *et al.* An Earth-sized planet with an Earth-like density. *Nature* **503,** 377–380 (2013).
47. Kopparapu, R. K. *et al*. Habitable zones around main-sequence stars: new estimates. *Astrophys. J*. **765,** 131 (2013).
48. Borucki, W. J. *et al*. Kepler-62: a five-planet system with planets of 1.4 and 1.6 Earth radii in the habitable zone. *Science* **340,** 587–590 (2013).
49. Borucki, W. J. *et al*. in *Transiting Extrasolar Planets Workshop* (eds Afonso, C., Weldrake, D. & Henning, T.) (Astro. Soc. Pacif. Conf., 2007).
50. Chaplin, W. J. & Miglio, A. Asteroseismology of solar-type and red-giant stars. *Annu. Rev. Astron. Astrophys*. **51,** 353–392 (2013).
51. Christensen-Dalsgaard, J. The new era of asteroseismology. *EAS Publications Series* **63,** 91–104 (2013).
52. Prsa, A. *et al.* Kepler eclipsing binary stars. I. Catalog and principal characterization of 1879 eclipsing binaries in the first data release. *Astron. J*. **141,** 83 (2011).
53. Tabachnik, S. & Tremaine, S. Maximum-likelihood method for estimating the mass and period distributions of extrasolar planets. *Mon. Not. R. Astron. Soc.* **335,** 151–158 (2002).
54. Youdin, A. N. The exoplanet census: a general method applied to Kepler. *Astrophys. J.* **742,** 38 (2011).
55. Howard, A. W. *et al*. Planet occurrence within 0.25 AU of solar-type stars from Kepler. *Astrophys. J. Suppl*. **201,** 15 (2012).
56. Dong, S. & Zhu, Z. Fast rise of 'Neptune-size' planets (4–8 $R_\oplus$) from P ~ 10 to ~250 days — statistics of Kepler planet candidates up to ~0.75 AU. *Astrophys. J*. **778,** 53 (2013).
57. Jenkins, J. M. *et al*. Overview of the Kepler science processing pipeline. *Astrophys. J.* **713,** L87–L91 (2010).
58. Tenenbaum, P. *et al*. Detection of potential transit signals in the first three quarters of Kepler mission data. *Astrophys. J. Suppl.* **199,** 24 (2012).

59. Christiansen, J. L. *et al*. Measuring transit signal recovery in the Kepler pipeline. I. Individual events. *Astrophys. J. Suppl.* **207,** 35 (2013).
60. Petigura, E. A. & Marcy, G. W. Identification and removal of noise modes in Kepler photometry. *Publ. Astron. Soc. Pacif.* **124,** 1073–1082 (2012).
61. Petigura, E. A., Howard, A. W. & Marcy, G. W. Prevalence of Earth-size planets orbiting Sun-like stars. *Proc. Natl Acad. Sci. USA* **110,** 19273–19278 (2013).
62. Petigura, E. A., Marcy, G. W. & Howard, A. W. A plateau in the planet population below twice the size of Earth. *Astrophys. J.* **770,** 69 (2013).
63. Batalha, N. Exploring exoplanet populations with NASA's Kepler mission. *Proc. Natl Acad. Sci. USA* http://dx.doi.org/10.1073/pnas.1304196111 (2014).
64. Wright, J. T. *et al*. The frequency of hot Jupiters orbiting nearby solar-type stars. *Astrophys. J.* **753,** 160 (2012).
65. Dawson, R. I. & Murray-Clay, R. A. Giant planets orbiting metal-rich stars show signatures of planet-planet interactions. *Astrophys. J.* **767,** L24 (2013).
66. Foreman-Mackey, D., Hogg, D. W. & Morton, T. D. Exoplanet population inference and the abundance of Earth analogs from noisy, incomplete catalogs. Preprint at http://arxiv.org/abs/1406.3020 (2014).
67. Lopez, E. D. & Fortney, J. J. Understanding the mass-radius relation for sub-neptunes: radius as a proxy for composition. *Astrophys. J.* **792,** 1 (2014).
68. Lithwick, Y., Xie, J. & Wu, Y. Extracting planet mass and eccentricity from TTV data. *Astrophys. J.* **761,** 122 (2012).
69. Hadden, S. & Lithwick, Y. Densities and eccentricities of 139 Kepler planets from transit time variations. *Astrophys. J.* **787,** 80 (2014).
70. Rogers, L. Most 1.6 Earth-radii planets are not rocky. Preprint at http://arxiv.org/abs/1407.4457 (2014).
71. Wu, Y. & Lithwick, Y. Density and eccentricity of Kepler planets. *Astrophys. J.* **772,** 74 (2013).
72. Weiss, L. M. & Marcy, G. W. The mass-radius relation for 65 exoplanets smaller than 4 Earth radii. *Astrophys. J.* **783,** L6 (2014).
73. Borucki, W. J. *et al*. Kepler-22b: a 2.4 Earth-radius planet in the habitable zone of a Sun-like star. *Astrophys. J.* **745,** 120 (2012).
74. Morton, T. D. & Swift, J. The radius distribution of small planets around cool stars. *Astrophys. J.* **791,** 10 (2014).
75. Lissauer, J. J. *et al*. Almost all of Kepler's multiple-planet candidates are planets. *Astrophys. J.* **750,** 112 (2012).
76. Fabrycky, D. C. *et al*. Architecture of Kepler's multi-transiting systems: II. New investigations with twice as many candidates. *Astrophys. J.* **790,** 146 (2014). **This article characterizes the orbital properties of Kepler's multi-planet systems.**
77. Tremaine, S. & Dong, S. The statistics of multi-planet systems. *Astron. J.* **143,** 94 (2012).
78. Dawson, R. I. *et al.* Large eccentricity, low mutual inclination: the three-dimensional architecture of a hierarchical system of giant planets. *Astrophys. J.* **791,** 89 (2014).
79. Ida, S. Stirring and dynamical friction rates of planetesimals in the solar gravitational field. *Icarus* **88,** 129–145 (1990).
80. Moorhead, A. V. *et al.* The distribution of transit durations for Kepler planet candidates and implications for their orbital eccentricities. *Astrophys. J. Suppl.* **197,** 1 (2011).
81. Plavchan, P., Bilinski, C. & Currie, T. Investigation of Kepler objects of interest stellar parameters from observed transit durations. *Publ. Astron. Soc. Pacif.* **126,** 34–47 (2014).
82. Kipping, D. M. Characterizing distant worlds with asterodensity profiling. *Mon. Not. R. Astron. Soc.* **440,** 2164–2184 (2014).
83. Kipping, D. M., Dunn, W. R., Jasinski, J. M. & Manthri, V. P. A novel method to photometrically constrain orbital eccentricities: multibody asterodensity profiling. *Mon. Not. R. Astron. Soc.* **421,** 1166–1188 (2012).
84. Dawson, R. I. & Johnson, J. A. The photoeccentric effect and proto-hot Jupiters. I. Measuring photometric eccentricities of individual transiting planets. *Astrophys. J.* **756,** 122 (2012).
85. Zakamska, N. L., Pan, M. & Ford, E. B. Observational biases in determining extrasolar planet eccentricities in single-planet systems. *Mon. Not. R. Astron. Soc.* **410,** 1895–1910 (2011).
86. Rafikov, R. R. & Slepian, Z. S. Dynamical evolution of thin dispersion-dominated planetesimal disks. *Astron. J.* **139,** 565–579 (2010).
87. Albrecht, S. *et al*. Obliquities of hot Jupiter host stars: evidence for tidal interactions and primordial misalignments. *Astrophys. J.* **757,** 18 (2012).
88. Fabrycky, D. & Tremaine, S. Shrinking binary and planetary orbits by Kozai cycles with tidal friction. *Astrophys. J.* **669,** 1298–1315 (2007).
89. Tremaine, S. On the origin of the obliquities of the outer planets. *Icarus* **89,** 85–92 (1991).
90. Heller, C. H. Encounters with protostellar disks. I. Disk tilt and the nonzero solar obliquity. *Astrophys. J.* **408,** 337–346 (1993).
91. Batygin, K. A primordial origin for misalignments between stellar spin axes and planetary orbits. *Nature* **491,** 418–420 (2012).
92. Sanchis-Ojeda, R. *et al*. Alignment of the stellar spin with the orbits of a three-planet system. *Nature* **487,** 449–453 (2012).
93. Hirano, T. *et al*. Measurements of stellar inclinations for Kepler planet candidates. II. Candidate spin-orbit misalignments in single- and multiple-transiting systems. *Astrophys. J.* **783,** 9 (2014).
94. Gladman, B. Dynamics of systems of two close planets. *Icarus* **106,** 247 (1993).
95. Smith, A. W. & Lissauer, J. J. Orbital stability of systems of closely-spaced planets. *Icarus* **201,** 381–394 (2009).
96. Lissauer, J. J. *et al*. Architecture and dynamics of Kepler's candidate multiple transiting planet systems. *Astrophys. J. Suppl.* **197,** 8 (2011).
97. Holman, M. J. A possible long-lived belt of objects between Uranus and Neptune. *Nature* **387,** 785–788 (1997).
98. Evans, N. W. & Tabachnik, S. A. Structure of possible long-lived asteroid belts. *Mon. Not. R. Astron. Soc.* **333,** L1–L5 (2002).
99. Fang, J. & Margot, J.-L. Are planetary systems filled to capacity? A study based on Kepler results. *Astrophys. J.* **767,** 115 (2013).
100. Eggenberger, A. in *EAS Publications Series* (eds Gozdziewski, K., Niedzielski, A. & Schneider, J.) Vol 42. 19–37 (2010).
101. Raghavan, D. *et al*. A survey of stellar families: multiplicity of solar-type stars. *Astrophys. J. Suppl.* **190,** 1–42 (2010).
102. Wang, J., Fischer, D. A., Xie, J.-W. & Ciardi, D. R. Influence of stellar multiplicity on planet formation. II. Planets are less common in multiple-star systems with separations smaller than 1500 AU. *Astrophys. J.* **791,** 111 (2014).
103. Tamuz, O. *et al*. The CORALIE survey for southern extra-solar planets. XV. Discovery of two eccentric planets orbiting HD 4113 and HD 156846. *Astron. Astrophys.* **480,** L33–L36 (2008).
104. Holman, M. J. & Wiegert, P. A. Long-term stability of planets in binary systems. *Astron. J.* **117,** 621–628 (1999).
105. Kipping, D. M. & Bakos, G. A., Buchhave, L., Nesvorny, D. & Schmitt, A. The hunt for exomoons with Kepler (HEK). I. Description of a new observational project. *Astrophys. J.* **750,** 115 (2012).
106. Kipping, D. M. et al. The hunt for exomoons with Kepler (HEK). II. Analysis of seven viable satellite-hosting planet candidates. *Astrophys. J.* **770,** 101 (2013).
107. Santos, N. C., Israelian, G. & Mayor, M. The metal-rich nature of stars with planets. *Astron. Astrophys.* **373,** 1019–1031 (2001).
108. Santos, N. C., Israelian, G. & Mayor, M. Spectroscopic [FeH] for 98 extra-solar planet-host stars. Exploring the probability of planet formation. *Astron. Astrophys.* **415,** 1153–1166 (2004).
109. Chiang, E. & Laughlin, G. The minimum-mass extrasolar nebula: *in situ* formation of close-in super-Earths. *Mon. Not. R. Astron. Soc.* **431,** 3444–3455 (2013).
110. Hansen, B. M. S. & Murray, N. Testing *in situ* assembly with the Kepler planet candidate sample. *Astrophys. J.* **775,** 53 (2013).
111. Rogers, L. A., Bodenheimer, P., Lissauer, J. J. & Seager, S. Formation and structure of low-density exo-Neptunes. *Astrophys. J.* **738,** 59 (2011).
112. Swift, J. J. *et al*. Characterizing the cool KOIs. IV. Kepler-32 as a prototype for the formation of compact planetary systems throughout the Galaxy. *Astrophys. J.* **764,** 105 (2013).
113. Goldreich, P. & Tremaine, S. Disk-satellite interactions. *Astrophys. J.* **241,** 425–441 (1980).
114. Baruteau, C. *et al*. Planet-disc interactions and early evolution of planetary systems. Preprint at http://arxiv.org/abs/1312.4293 (2013).
115. Peale, S. J. Orbital resonances in the Solar System. *Ann. Rev. Astron. Astrophys.* **14,** 215–246 (1976).
116. Rein, H. Period ratios in multiplanetary systems discovered by Kepler are consistent with planet migration. *Mon. Not. R. Astron. Soc.* **427,** L21–L24 (2012).
117. Goldreich, P. & Schlichting, H. E. Overstable librations can account for the paucity of mean motion resonances among exoplanet pairs. *Astron. J.* **147,** 32 (2014).
118. Petrovich, C., Tremaine, S. & Rafikov, R. Scattering outcomes of close-in planets: constraints on planet migration. *Astrophys. J.* **786,** 101 (2014).
119. Lithwick, Y. & Wu, Y. Resonant repulsion of Kepler planet pairs. *Astrophys. J.* **756,** L11 (2012).
120. Petrovich, C., Malhotra, R. & Tremaine, S. Planets near mean-motion resonances. *Astrophys. J.* **770,** 24 (2013).
121. Schlaufman, K. C. Tests of *in situ* formation scenarios for compact multiplanet systems. *Astrophys. J.* **790,** 91 (2014).
122. Smith, J. C. *et al*. Kepler presearch data conditioning II. A Bayesian approach to systematic error correction. *Publ. Astron. Soc. Pacif.* **124,** 1000–1014 (2012).
123. Stumpe, M. C. *et al*. Kepler presearch data conditioning I. Architecture and algorithms for error correction in Kepler light curves. *Publ. Astron. Soc. Pacif.* **124,** 985–999 (2012).
124. Steffen, J. H. *et al*. Kepler constraints on planets near hot Jupiters. *Proc. Natl Acad. Sci. USA* **109,** 7982–7987 (2012).
125. Dong, S. *et al.* On the metallicities of Kepler stars. *Astrophys. J.* **789,** L3 (2014).
126. Wright, J. T. *et al.* The exoplanet orbit database. *Publ. Astron. Soc. Pacif.* **123,** 412–422 (2011).
127. Latham, D. W. *et al.* A first comparison of Kepler planet candidates in single and multiple systems. *Astrophys. J.* **732,** L24 (2011).
128. Lopez, E. D., Fortney, J. J. & Miller, N. How thermal evolution and mass-loss sculpt populations of super-Earths and sub-Neptunes: application to the Kepler-11 system and beyond. *Astrophys. J.* **761,** 59 (2012).

Acknowledgments This research has made use of the Exoplanet Orbit Database at http://exoplanets.org and the Extrasolar Planets Encyclopedia at http://exoplanets.eu. We are grateful to the Kepler Science Team for their extensive efforts in producing the high-quality data set that has made possible the results reviewed here. We thank W. Borucki, E. Chiang, S. Dong, E. Lee, E. Lopez, L. Rogers, J. Rowe, A. Youdin and K. Zahnle for helpful discussions and comments on the manuscript. R.I.D. and S.T. gratefully acknowledge funding from the Miller Institute for Basic Research in Science at the University of California, Berkeley.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this paper at go.nature.com/fimnn9. Correspondence should be addressed to J. L. (jack.lissauer@nasa.gov).

REVIEW

doi:10.1038/nature13782

Highlights in the study of exoplanet atmospheres

Adam S. Burrows[1]

Exoplanets are now being discovered in profusion. To understand their character, however, we require spectral models and data. These elements of remote sensing can yield temperatures, compositions and even weather patterns, but only if significant improvements in both the parameter retrieval process and measurements are made. Despite heroic efforts to garner constraining data on exoplanet atmospheres and dynamics, reliable interpretation has frequently lagged behind ambition. I summarize the most productive, and at times novel, methods used to probe exoplanet atmospheres; highlight some of the most interesting results obtained; and suggest various broad theoretical topics in which further work could pay significant dividends.

The modern era of exoplanet research started in 1995 with the discovery of the planet 51 Pegasi b[1], when astronomers detected the periodic radial-velocity Doppler wobble in its star, 51 Peg, induced by the planet's nearly circular orbit. With these data, and knowledge of the star, the orbital period (P) and semi-major axis (a) could be derived, and the planet's mass constrained. However, the inclination of the planet's orbit was unknown and, therefore, only a lower limit to its mass could be determined. With a lower limit of 0.47 M_J (where M_J is the mass of Jupiter) and given its proximity to its primary (a is about 0.052 AU, 1 AU being the Earth–Sun distance; one hundred times closer to its star than Jupiter is to the Sun), the induced Doppler wobble is optimal for detection by the radial-velocity technique. The question was how such a 'hot Jupiter' could exist and survive. Although its survival is now understood (see 'Winds from planets'), the reason for its close orbital position is still a subject of vigorous debate. Nevertheless, such close-in giants are selected for using the radial-velocity technique and soon scores, then hundreds, of such gas giants were discovered in this manner.

However, aside from a limit on planet mass, and the inference that proximity to its star leads to a hot (1,000–2,000 kelvin (K)) irradiated atmosphere, no useful physical information on such planets was available with which to study planet structure, their atmospheres or composition. A breakthrough along the path to characterization, and the establishment of mature exoplanet science, occurred with the discovery of giant planets, still close-in, that transit the disk of their parent star. The chance of a transit is larger if the planet is close, and HD 209458b, which is about 0.05 AU from its star, was the first to be found[2]. Optical measurements yielded a radius for HD 209458b of about 1.36 R_J, where R_J is the radius of Jupiter. Jupiter is roughly ten times, and Neptune is roughly four times, the radius of Earth (R_E). Since then, hundreds of transiting giants have been discovered using ground-based facilities. The magnitude of the attendant diminution of a star's light during such a primary transit (eclipse) by a planet is the ratio of their areas (the transit depth, R_p^2/R_*^2, where R_p is the planet's radius and R_* is the star's radius), so with knowledge of the star's radius, the planet's radius can be determined. Along with radial-velocity data, because the orbital inclination of a planet in transit is known, one then has a radius–mass pair with which to do some science. The transit depth of a giant passing in front of a Sun-like star is about 1%, and such a large magnitude can easily be measured with small telescopes from the ground. A smaller, Earth-like planet requires the ability to measure transit depths 100 times more precisely. It was not long before many hundreds of gas giants were detected both in transit and by the radial-velocity method, the former requiring modest equipment and the latter requiring larger telescopes with state-of-the-art spectrometers with which to measure the small stellar wobbles. Both techniques favour close-in giants, so for many years these objects dominated the bestiary of known exoplanets.

Better photometric precision near or below one part in 10^4–10^5, which is achievable only from space, is necessary to detect the transits of Earth-like and Neptune-like exoplanets across Sun-like stars, and, with the advent of Kepler[3] and CoRoT (Convection, Rotation and Planetary Transits)[4], astronomers have now discovered a few thousand exoplanet candidates. Kepler in particular revealed that most planets are smaller than about 2.5 R_E (four times smaller than Jupiter), but fewer than around 100 of the Kepler candidates are close enough to us to be measured with state-of-the-art radial-velocity techniques. Without masses, structural and bulk compositional inferences are problematic. Moreover, most of these finds are too distant for photometric or spectroscopic follow-up from the ground or space to provide thermal and compositional information.

A handful of the Kepler and CoRoT exoplanets, and many of the transiting giants and 'sub-Neptunes' discovered using ground-based techniques, are not very distant and have been photometrically and spectroscopically followed up using both ground-based and space-based assets to help to constrain their atmospheric properties. In this way, and with enough photons, some information on atmospheric compositions and temperatures has been revealed for around 50 exoplanets, mostly giants. However, even these data are often sparse and ambiguous, rendering most such hard-won results provisional[5]. The nearby systems hosting larger transiting planets around smaller stars are the best targets for a programme of remote sensing to be undertaken, but such systems are a small subset of the thousands of exoplanets currently in the catalogues.

One method by which astronomers are performing such studies is by measuring the transit radius as a function of wavelength[6–8]. Because the opacity of molecules and atoms in a planet's atmosphere is a function of wavelength, the apparent size of the planet is also a function of wavelength — in a manner that is characteristic of atmospheric composition. Such a 'radius spectrum' can reveal the atmosphere's composition near the planet terminators, but the magnitude of the associated variation is

[1]Astrophysical Sciences, Princeton University, 4 Ivy Lane, Princeton, New Jersey 08544, USA.

down from the average transit depth by a factor of around $2H/R_p$, where H is the atmospheric scale height (a function of average temperature and gravity). This ratio can be ~0.1 to 0.01, correspondingly making it more difficult to determine a transit radius spectrum. Only telescopes such as the Spitzer Space Telescope[9], the Hubble Space Telescope and the largest ground-based telescopes with advanced spectrometers are up to the task, and even then the results can be difficult to interpret.

Another method probes the atmospheres of transiting exoplanets at secondary eclipse, when the star occults the planet about 180° out of phase with the primary transit. The abrupt difference between the summed spectrum of planet and star just before and during the eclipse of the planet by the star is the planet's spectrum at full face. Secondary eclipse spectra include reflected (mostly in the optical and near-ultraviolet region) and thermally emitted (mostly in the near- and mid-infrared region) light, and models are necessary to distinguish, if possible, the two components. It should be noted that separate images of the planet and star are not obtained by this technique, and a planet must be transiting. With few exceptions, when the planet does not transit, the summed light of a planet and star varies too slowly and smoothly for such a variation to be easily distinguished from the systematic uncertainties of the instruments to reveal the planet's emissions as a function of orbital phase. For the close-in transiting hot Jupiters, the planet flux in the near-infrared is 10^{-3} times the stellar flux — much higher than the ratio expected for the class of planet in a wide orbit that can be separated from its primary star by high-contrast imaging techniques (see 'High-contrast imaging'). In cases when such high-contrast direct imaging is feasible, the planet is farther away from the star (hence, dim) and difficult to discern from under the stellar glare. However, hot, young giants can be self-luminous enough to be captured by current high-contrast imaging techniques, and a handful of young giant planets have been discovered and characterized by this technique. More are expected as the technology matures[10–13].

The secondary eclipse and primary transit methods used to determine or constrain atmospheric compositions and temperatures (as well as other properties) generally involve low-resolution spectra with large systematic and statistical errors. These methods are complementary in that transit spectra reliably reveal the presence of molecular and atomic features, and are an indirect measure of temperature through the pressure-scale height, whereas the flux levels of secondary eclipse spectra scale directly with temperature, but could in fact be featureless for an isothermal atmosphere. The theoretical spectra with which they are compared in order to extract parameter values are also imperfect, and this results in less trustworthy information than one would like. Giant planets (and 'Neptunes') orbiting closely around nearby stars are the easiest targets, and are the stepping stones to 'Earths'. Secondary and primary transit spectral measurements of Earth-like planets around Sun-like stars, as well as direct high-contrast imaging of such small planets, are not currently feasible. However, measurements of exo-Earths around smaller M-dwarf stars might be, if suitable systems can be found. Nevertheless, with a few score of transit and secondary eclipse spectra, some planetary phase light curves, a few high-contrast campaigns and measurements, and some narrow-band, but very high spectral resolution measurements using large telescopes, the first generation of exoplanet-atmosphere studies has begun.

There are several helpful reviews on the theory of exoplanet atmospheres[14–22]. Added to these, there are informed discussions on the molecular spectroscopy and opacities that are central to model building[23–27]. Monographs on the relevant thermochemistry and abundances have been published over the years[28–32]. In this Review, I do not attempt to cover the literature of detections and claims, nor do I attempt to review the thermochemical, spectroscopic or dynamical modelling efforts so far. Instead, I focus on those few results concerning exoplanet atmospheres that to my mind stand out, that seem most robust and that collectively summarize what we have truly learned. I present, of necessity, only a small subset of the published literature, and no doubt some compelling results have been neglected for lack of space. In addition, I touch on only the basics of the atmosphere theory applied so far, preferring to focus, when possible, on the progress in theory that is necessary for the next generation of exoplanet-atmosphere studies to evolve productively. I embark on a discussion of what I deem to be a few of the milestone observational papers in core topics; these might be considered to constitute the spine of progress in recent exoplanet-atmosphere study. I accompany each with a short discussion of the associated theoretical challenges posed by the data.

Transit detection of atoms and molecules

The apparent transit radius of a planet with a gaseous atmosphere is the impact parameter of a ray of stellar light for which the optical depth at that wavelength (λ) is of order unity. It should be noted that at that level the corresponding radial optical depth, which if in absorption is relevant to emission spectra at secondary eclipse, will be much smaller. Because an atmosphere has a thickness (extent), and because absorption and scattering cross-sections are functions of photon wavelength that in combination with the air column constitute optical depth, the measured transit radius is a function of wavelength. Therefore, measurements of a planet's transit depths at many wavelengths of light reveal its atomic and molecular composition. A good approximation for this is given by[33]:

$$\mathrm{d}R_p/\mathrm{dln}(\lambda) \sim H\ \mathrm{dln}\,\sigma(\lambda)/\mathrm{dln}(\lambda) \qquad (1)$$

where σ(λ) is the composition-weighted total cross-section and the scale height, H, is $kT/\mu g$, where g is the planet's surface gravity, μ is the mean molecular weight, T is an average atmospheric temperature, and k is Boltzmann's constant. H sets the scale of the magnitude of potential fluctuations of R_p with λ, and σ(λ) is determined mostly by the atomic and molecular species in the atmosphere.

Charbonneau *et al.*[34] were the first to successfully use this technique with the 4-σ measurement of atomic sodium in the atmosphere of HD 209458b. Along with HD 189733b, this nearby giant planet has been the most photometrically and spectroscopically scrutinized. Since then, Sing *et al.*[35] have detected potassium in XO-2b and Pont *et al.*[36,37] have detected both sodium and potassium in HD 189733b. These are all optical measurements at and around the sodium D doublet (about 0.589 μm) and the potassium resonance doublet (around 0.77 μm), and reveal the telltale differential transit depths in and out of the associated lines.

Based on the study of brown dwarfs, the presence of neutral alkali metals in the atmospheres of irradiated exoplanets with similar atmospheric temperatures (~1,000–1,500 K) was expected, and their detection was gratifying. Indeed, there is a qualitative correspondence between the atmospheres of close-in and irradiated, or young giant planets (with masses of order M_J) and older brown dwarfs (with masses of tens of M_J). Alkalis persist to lower temperatures (~800–1,000 K) and are revealed in close-in exoplanet transit and emission spectra, and in older brown-dwarf emission spectra because silicon and aluminium, with which they would otherwise combine to form feldspars, are sequestered at higher temperatures and greater depths into more refractory species, and rained out. Had the elements with which sodium and potassium would have combined persisted in the atmosphere at altitude, these alkalis would have combined and their atomic form would not have been detected[38]. The more refractory silicates (and condensed iron) reside in giant exoplanets (and in Jupiter and Saturn), but at great depths. In L-dwarf brown dwarfs, they are at the surface, reddening the emergent spectra significantly.

However, the strength, in transiting giant exoplanets, of the contrast in and out of these atomic alkali lines is generally less than expected[8]. Subsolar elemental sodium and potassium abundances, ionization by stellar light, and hazes have been invoked to explain the diminished strength of their associated lines, but the haze hypothesis is gaining ground. The definition of a haze can merge with that of a cloud, but generally hazes are clouds of small particulates at altitude that may be condensates of trace species or products of photolysis by stellar

ultraviolet light and polymerization. They are generally not condensates of common or abundant molecular species (such as water, ammonia, iron or silicates, none of which fits the bill here). Although it is not at all clear what this haze is, hazes at altitude (<0.01 bars) can provide a nearly featureless continuum opacity to light and easily mute atomic and molecular line strengths. Indeed, hazes are emerging as central and ubiquitous features in exoplanet atmospheres. Annoyingly, not much mass is necessary to have an effect on transit spectra, making quantitative interpretation all the more difficult. The fact that the red colour of Jupiter itself is produced by a trace species (perhaps a haze) that so far has not been identified is a sobering testament to the difficulties that lie ahead in completely determining exoplanet atmospheric compositions.

The multi-frequency transit measurements of HD 189733b from the near-ultraviolet to the mid-infrared by Pont *et al.*[36,37] are the clearest and most marked indications that some exoplanets have haze layers (Fig. 1). Curiously, the measurements show no water or other molecular features in transit. Aside from the aforementioned sodium and potassium atomic features in the optical, the transit spectrum of HD 189733b is consistent with a featureless continuum. Water features in a hydrogen (H_2) atmosphere are very difficult to completely suppress, so their absence is strange. Furthermore, the transit radius increases below about 1.0 μm with decreasing wavelength in a manner that is reminiscent of Rayleigh scattering. However, owing to the large cross-sections implied, the culprit can only be a haze or a cloud. It should be mentioned that these transit data cannot distinguish between absorption and scattering, although scattering is the more likely cause for most plausible haze materials and particle sizes. Scattering is also indicated by the near lack of evidence for absorbing particulates in HD 189733b secondary eclipse emission spectra[39]. Together, these data suggest that a scattering haze layer at altitude is obscuring the otherwise distinctive spectral features of the spectroscopically active atmospheric constituents.

Transit spectra for the mini-Neptune GJ 1214b have been taken by many groups, but the results concerning possible distinguishing spectral features have, until recently, been quite ambiguous[40]. In principle, there are diagnostic water features at around 1.15 μm and 1.4 μm. However, Kreidberg *et al.*[41], using the Wide Field Camera-3 (WFC3) on the Hubble Space Telescope, have demonstrated that from ~1.1 to 1.6 μm its transit spectrum is around 5–10 times flatter than a water-rich, H_2-dominated atmosphere with a solar abundance of water (oxygen) (Fig. 2). Flatness could indicate that the atmosphere has no scale height (see equation 1) (for example, due to a high mean molecular weight, μ), or herald the presence, yet again, of a thick haze layer obscuring the molecular features. Not surprisingly, a panchromatic obscuring haze layer is currently the front runner.

Lest one think that hazes completely mask the molecules of exoplanet atmospheres, Deming *et al.*[42] have published transit spectra of HD 209458b (Fig. 3) and XO-1b that clearly show the water feature at around 1.4 μm. However, the expected accompanying water feature at about 1.15 μm is absent. The best interpretation of this is that this feature is suppressed by the presence of a haze with a continuum, although wavelength-dependent, interaction cross-section that trails off at longer wavelengths. The weaker apparent degree of suppression in these exoplanet atmospheres might suggest that their hazes are thinner or deeper (at higher pressures) than in HD 189733b. Physical models explaining this behaviour are lacking.

So, the only atmospheric species that have clearly been identified in transit are water, sodium, potassium and a 'haze'. Molecular hydrogen is the only gas with a low enough μ to provide a scale height that is great enough to explain the detection in transit of any molecular features (see equation 1) in a hot, irradiated atmosphere, and I would include it as indirectly indicated. However, carbon monoxide, carbon dioxide, ammonia, nitrogen gas, acetylene, ethylene, phosphine, hydrogen sulphide, oxygen, ozone, nitrous oxide and hydrogen cyanide have all been proferred as exoplanet atmosphere gases. Clearly, the field is in its spectroscopic infancy. Facilities such as

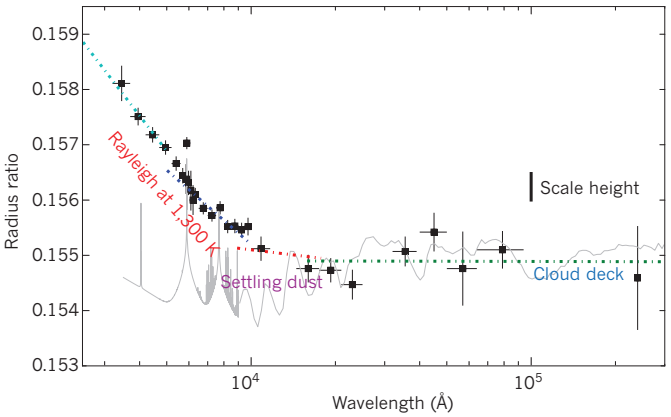


Figure 1 | Transit spectrum of giant exoplanet HD 189733b. The planet/star radius ratio against wavelength in ångströms. The black dots are the data points and the dotted lines are models. From left to right the dotted lines show the possible effect of Rayleigh scattering by mixed small grains at 2,000 K and at 1,300 K, by settling grains and by an opaque cloud deck. The grey line is an example spectrum without a haze. Reprinted with permission from ref. 36.

next-generation ground-based telescopes (extremely large telescopes, ELTs) and space-based telescopes such as the James Webb Space Telescope (JWST)[22], or a dedicated exoplanet space-based spectrometer, will be essential if transit spectroscopy is to realize its true potential for exoplanet atmospheric characterization. The JWST in particular will have spectroscopic capability from ~0.6 to ~28.3 μm and will be sensitive to most of the useful atmospheric features expected in giant, Neptune-like and sub-Neptune exoplanets. It may also be able to detect and characterize a close-in Earth or super-Earth around a nearby small M star.

There are a number of theoretical challenges that must be met before transit data can be converted into reliable knowledge. Such spectra probe the terminator region of the planet that separates the day and night sides. They sample the transitional region between the hotter day and cooler night of the planet, at which the compositions may be changing and condensates may be forming. Hence, the compositions extracted may not be representative even of the bulk atmosphere. Ideally, one would want to construct dynamical three-dimensional (3D) atmospheric circulation models that couple non-equilibrium chemistry and detailed molecular opacity databases with multi-angle 3D radiation transfer. Given the emergence of hazes and clouds as potentially important features of exoplanet atmospheres, a meteorologically credible condensate model is also desired. We are far from the latter[43], and the former's capabilities are only now being constructed, with limited success[44]. The dependence of transit spectra on species abundance is weak, making it now difficult to derive mixing ratios from transit spectra to better than a factor of 10 to 100. Although the magnitude of the variation of apparent radius with wavelength depends on atmospheric scale height, and hence temperature, the temperature–pressure profile and the variation of abundance with altitude are not easily constrained. To obtain even zeroth-order information, one frequently creates isothermal atmospheres with chemical equilibrium or uniform composition. Current haze models are ad hoc, and adjusted a posteriori to fit the all-too-sparse and at times ambiguous data. To justify the effort necessary to do better will require much improved and higher-resolution measured spectra[5].

Data at secondary eclipse require a similar modelling effort, but probe the integrated flux of the entire dayside. Hence, a model that correctly incorporates the effects of stellar irradiation ('instellation') and limb effects is necessary. Moreover, the flux from the cooling planetary core, its longitudinal and latitudinal variation, and a circulation model that redistributes energy and composition are needed. Most models employed so far use a representative one-dimensional (planar) approximation, and radiative and chemical equilibrium for what is a

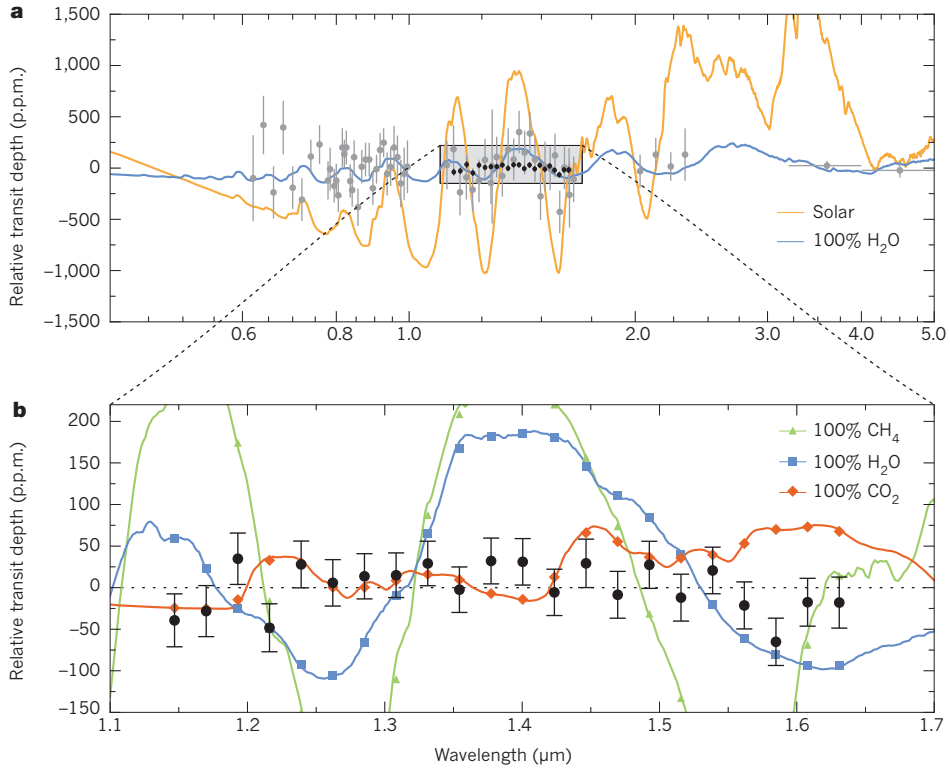


Figure 2 | The transmission spectrum of GJ 1214b. The relative depth of the transit of the sub-Neptune GJ 1214b against wavelength (1.1 μm to 1.7 μm). The coloured lines in both (a) and (b) are various transit spectral models without a haze (cloud-free solar composition, orange; 100% water, blue; methane, green; and carbon dioxide, red). The data (black and grey dots) are effectively flat, ruling out all models shown and suggesting a veiling haze.

hemispherical region that might be out of chemical equilibrium (and slightly out of radiative equilibrium). The emission spectra of the day-side depend more on the absorptive opacities, whereas transit spectra depend on both scattering and absorption opacities. Hence, if the haze inferred in some transit spectra is due predominantly to scattering, its effect on secondary eclipse spectra will be minimal, making it slightly more difficult to use insight gained from one to inform the modelling of the other.

Many giant exoplanets, and a few sub-Neptunes, have been observed at secondary eclipse, but the vast bulk of these data are comprised of a few photometric points per planet. The lion's share have been garnered using Spitzer, Hubble or large-aperture ground-based telescopes, and pioneering attempts to inaugurate this science were carried out by Deming *et al.*[45] and Charbonneau *et al.*[46]. Photometry, particularly if derived using techniques that are subject to systematic errors, is ill-suited to delivering solid information on composition, thermal profiles or atmospheric dynamics. The most one can do with photometry at secondary eclipse is to determine rough average emission temperatures, and perhaps reflection albedos in the optical. Temperatures for close-in giant exoplanet atmospheres from around 1,000 to 3,000 K have, in this way, been determined. Of course, the mere detection of an exoplanet is a victory, and the efforts that have gone into winning these data should not be discounted. Nevertheless, with nearly 50 such campaigns and detections 'in the can', we have learned that it is only with next-generation spectra that use improved (perhaps dedicated) spectroscopic capabilities that the desired thermal and compositional information will be forthcoming.

One of the few reliable compositional determinations at secondary eclipse obtained so far is for the dayside atmosphere of HD 189733b using the now-defunct infrared spectrograph on-board Spitzer[39]. This very-low-resolution spectrum nevertheless provided a 3σ detection of water at around 6.2 μm. Other papers have reported the detection of molecules at secondary eclipse, but many are less compelling, and earlier reports of water being detected using photometry alone at secondary eclipse are very model-dependent[47]. It is only with well-calibrated spectra that one can determine with confidence the presence in any exoplanet atmosphere of any molecule or atom.

Winds from planets

The existence of what are now somewhat contradictorily called hot Jupiters has, since the discovery of 51 Peg b in 1995, been somewhat of a puzzle. These planets probably cannot form as close as they are observed to their parent star and must migrate in, by some process, from beyond the so-called ice line. In such cold regions, ices can form and accumulate to nucleate gas-giant formation. Subsequent inward migration could be driven early in the planet's life by gravitational torquing by the proto-planetary disk or by planet–planet scattering, followed by tidal dissipation in the planet (which circularizes its orbit). However, once parked at between ~0.01 AU and 0.1 AU from the star, how does the gaseous planet, or a gaseous atmosphere of a smaller planet, survive evaporation by the star's intense irradiation during perhaps billions of years seemingly in extremis? The answer is that for sub-Neptunes and rocky planets their atmospheres or gaseous envelopes might not survive, but for more massive gas giants the gravitational well at their surfaces may be sufficiently deep. Nevertheless, since the first discoveries, evaporation has been of interest[48]. The atmospheres of Earth and Jupiter are known to be evaporating, although at a very low rate. But what happens to a hot Jupiter that experiences 10^4 times the instellation that Jupiter does?

The answer came with the detection by Vidal-Madjar *et al.*[49] of a wind from HD 209458b. Using the transit method, but in the ultraviolet around the Lyman-α line of atomic hydrogen at around 0.12 μm, the authors measured a transit depth of about 15%. Such a large depth implies a planet radius greater than four R_J, which is not only much greater than what is inferred in the optical, but beyond the tidal Roche radius. Matter at such distances is not bound to the planet, and the only plausible explanation was that a wind was being blown off the planet. The absorption cross-sections in the ultraviolet are huge, so the matter densities that are necessary to generate a transverse chord optical depth

of one are very low — too low to affect the optical and infrared measurements. The upshot of this is the presence of a quasi-steady planetary wind with a mass-loss rate of 10^{10}–10^{11} gm s^{-1}. At that rate, HD 209458b will lose no more than around 10% of its mass in Hubble time.

Since this initial discovery, winds from the hot Jupiters HD 189733b[50] and WASP-12b[51], and from the hot Neptune GJ 436b[52] have been discovered by the ultraviolet transit method and partially characterized. In all cases, the telltale indicator was in atomic hydrogen. Mass-loss rates have been estimated[53], and in the case of WASP-12b might be sufficient to completely evaporate the giant within as little as about 1 gigayear. The presence of atomic hydrogen implies the photolytic or thermal break-up of molecular hydrogen, so these data simultaneously suggest the presence of both H and H_2. Linsky *et al.*[54] detected ionized carbon and silicon in HD 209458b's wind, and Fossati *et al.*[51] detected ionized magnesium in WASP-12b's wind, but the interpretation of the various ionized species detected in these transit-observation campaigns is ongoing.

The theoretical challenges posed by planetary winds revolve partly around the driver. Is the wind driven by the subset of the instellation represented by the ultraviolet and X-ray component of the total stellar flux? In addition, in the rotating system of the orbiting planet, what ingress or egress asymmetries in the morphology of the wind exist? There are indications that Coriolis forces on planet winds are indeed shifting the times of ingress and egress. What is the effect of planet–star wind interactions? There are suggestions of Doppler shifts in lines of the ultraviolet transit data that arise from planet-wind speeds, but how can we be sure? How is the material for the wind replenished from the planet atmosphere and interior? And finally, what is the correspondence between the ultraviolet photolytic chemistry in the upper reaches of the atmosphere that modifies its composition there and wind dynamics? This is a rich subject tied to many subfields of science, and is one of the important topics to emerge from transit spectroscopy.

Phase light curves and planet maps

As a planet traverses its orbit, its brightness, as measured at Earth at a given wavelength, varies with orbital phase. A phase light curve comprises both a reflected component that is a stiff function of the star–planet–Earth angle and is most prominent in the optical and ultraviolet; and a thermal component that more directly depends on the temperature and composition of the planet's atmosphere, and their longitudinal variation around the planet, and is most prominent in the near- and mid-infrared. Hence, a phase light curve is sensitive to the day–night contrast and is a useful probe of planetary atmospheres[55–59]. It should be mentioned that the planet/star contrast ratio is largest for large exoplanets in the closest orbits, so hot Jupiters currently provide the best targets.

In the optical, there has been some work to derive the albedo[55,56], or reflectivity, of close-in exoplanets, which is largest when there are reflecting clouds and smallest when the atmosphere is absorbing. In the latter case, thermal emission at high atmospheric temperatures can be mistaken for reflection, so detailed modelling is required. In any case, Kepler, with its superb photometric sensitivity, has been used to determine optical phase curves[60] of a few exo-giants in the Kepler field, and the MOST (Microvariability and Oscillations of Stars) microsatellite has put a low upper limit on the optical albedo of HD 209458b[61,62], but much remains to be done to extract diagnostic optical phase curves and albedos for exoplanets.

Interesting progress has been made, however, in the thermal infrared. Using Spitzer at 8 μm, Knutson *et al.*[63] not only derived a phase light curve for HD 189733b, but derived a crude thermal map of its surface. By assuming that the thermal emission pattern over the planet surface was fixed during the observations, they derived the day–night brightness contrast (translated into a brightness temperature at 8 μm) and a longitudinal brightness temperature distribution. In particular, they measured the position of the 'hot spot'. If the planet is in synchronous rotation (spin period is the same as the orbital period), and there are no equatorial winds to advect heat around the planet, one would expect the hot spot to be at the substellar point. The light curve would phase up with the orbit and the peak brightness would occur at the centre of secondary eclipse. However, what the authors observed was a shift downwind to the east by around 16° ± 6. The most straightforward interpretation is that the stellar heat absorbed by the planet is advected downstream before being re-radiated by super-rotational flows such as those that are observed on Jupiter itself. Moreover, these data indicate that, because the measured day–night brightness temperature contrast was only about 240 K, the zonal wind flows driven by stellar irradiation carry heat to the night side, where it is radiated at a detectable level. Hence, these data point to the existence of atmospheric dynamics on HD 189733b, qualitatively (although not quantitatively) in line with theoretical expectations[44].

For HD 189733b, this work has been followed up using Spitzer at 3.6 μm and 4.5 μm[64] and, in a competing effort, a more refined map has been produced[65]. Infrared phase curves for the giants HD 149026b[66], HAT-P-2b[67] and WASP-12b[68], among other exoplanets, have been obtained. However, one of the most intriguing phase curves was obtained by Crossfield *et al.*[69] using Spitzer at 24 μm for the non-transiting planet υ Andromedae b (Fig. 4). The authors found a huge phase offset of around 80°, for which a cogent explanation is still lacking. The closeness of this planet to Earth could partly compensate for the fact that it is not transiting to allow sufficient photometric accuracy without eclipse calibration, yielding one of the few non-transiting light curves. All these efforts collectively demonstrate the multiple, at times unanticipated and creative, methods being employed by observers seeking to squeeze whatever information they can from exoplanets.

Theoretical models for light curves have been sophisticated, but theory and measurement have not yet meshed well. Both need to be improved. First, models need to be improved in terms of their treatment of hazes and clouds that could reside in exoplanet atmospheres and will boost reflection albedos significantly; second, they need to incorporate polarization to realize its diagnostic potential[59,70]; third, they should constrain the possible range of phase functions to aid in retrievals; fourth, they need to embed the effects of variations in planet latitude and longitude in the analysis protocols; fifth, they should provide observational diagnostics with which to probe atmospheric pressure depths, particularly using multi-frequency data; sixth, they should be constructed as a function of orbital eccentricity, semi-major axis,

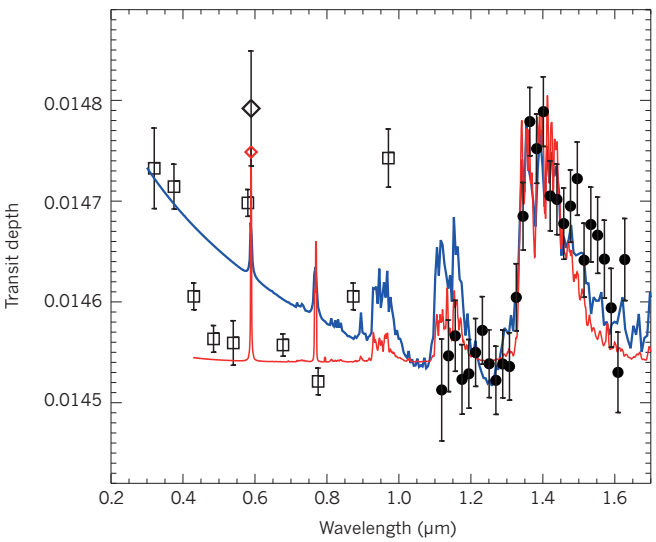


Figure 3 | Transit depth spectrum of the hot Jupiter HD 209458b. Data points are shown as black circles and open squares/diamonds. The presence of water is demonstrated by the occurrence of a feature at 1.4 μm, but the corresponding ~1.15-μm feature is absent. The best explanation is that the latter is suppressed by haze scattering. Not obvious here is the fact that even the 1.4-μm feature is muted with respect to non-haze models. The two coloured curves are representative model spectra with different levels of haze. Reprinted with permission from ref. 42.

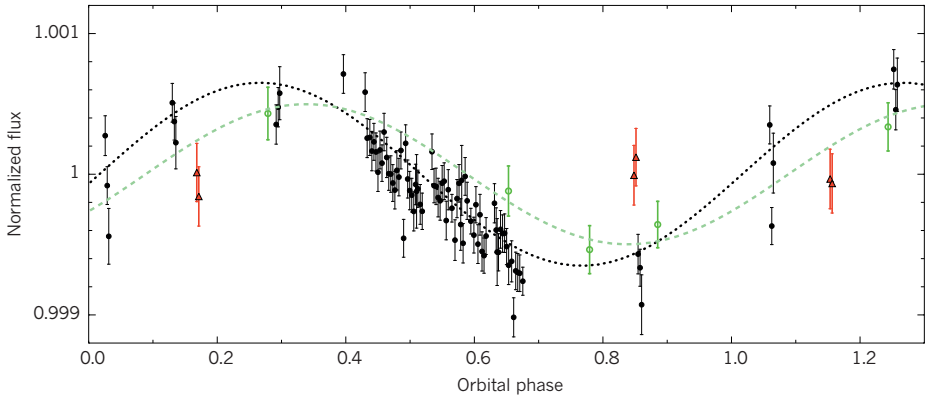


Figure 4 | The measured light curve of υ Andromedae b. The thin black dotted curve is the authors' best fit to the data points (black dots with error bars at a wavelength of 24 μm), showing a phase offset of ~80° (22% of a circuit). Reprinted with permission from ref. 69.

and inclination; and finally they should span the wide range of masses and compositions that the heterogeneous class of exoplanets is likely to occupy. Accurate spectral data with good time coverage from the optical to the mid-infrared could be game-changing, but theory needs to be ready with useful physical diagnostics.

High spectral resolution techniques

The intrinsic dimness of planets under the glare of stars renders high-resolution, panchromatic spectral measurements difficult, if desirable. However, ultra-high spectral resolution measurements using large-aperture ground-based telescopes, but over a very narrow spectral range and targeting molecular band features in a planet's atmosphere that are otherwise jumbled together at lower resolutions, has recently been demonstrated. Snellen *et al.*[71] have detected the Doppler variation owing to HD 209458b's orbital motion of carbon monoxide features near ~2.3 μm. The required spectral resolution ($\lambda/\Delta\lambda$) was about 10^5 and the plant's projected radial velocity just before and just after primary transit changed from +15 km s^{-1} to −15 km s^{-1}. This is consistent with the expected circular orbital speed of around 140 km s^{-1} and provides an unambiguous detection of carbon monoxide. Furthermore, this team attempted to measure the zonal wind speeds of air around the planet, estimated theoretically to be near ~1 km s^{-1}, thereby demonstrating the potential of such a novel technique to extract weather features on giant exoplanets. The same basic method has been applied near primary transit to detect carbon monoxide[72] and water[73] in HD 189733b. Carbon monoxide can be detected in Jupiter and was thermochemically predicted to exist in abundance in the atmospheres of hot Jupiters[31], but its actual detection by this method is impressive.

In fact, the same technique has been successfully applied in the carbon monoxide band to the non-transiting planet τ Boötis b[74] and for the wide-separation giant planet (or brown dwarf) β Pictoris b[75], verifying the presence of carbon monoxide in both their atmospheres. Finally, using a related technique Crossfield *et al.*[76] have been able to conduct high-resolution 'Doppler imaging' of the closest known brown dwarf (Luhman 16B). By assuming that the brown dwarf's surface features are frozen during the observations and that it is in solid-body rotation, and by dividing the surface into a grid in latitude and longitude, they were able to determine (by model fitting) surface brightness variations from the variations of its flux and Doppler-shift time series. By this means, they mapped surface spotting that may reflect broken cloud structures (Fig. 5).

In support of such measurements, theory needs to refine its modelling of planet surfaces, zonal flows and weather features, 3D heat redistribution and velocity fields, and temporal variability. Currently, most 3D general circulation models do not properly treat high Mach number flows, but they predict zonal wind Mach numbers of order unity. There are suggestions that magnetic fields affect the wind dynamics and heating in the atmosphere, but self-consistent multi-dimensional radiation magnetohydrodynamic models have not yet been constructed.

This series of measurements of giant exoplanets and brown dwarfs using high-resolution spectroscopy focused on narrow molecular features emphasizes two important aspects of exoplanet research. The first is that observers can be clever and develop methods unanticipated in roadmap documents and decadal surveys. The second is that with the next-generation of ground-based ELTs equipped with impressive spectrometers, astronomers may be able to measure and map some exoplanets without using the high-contrast imaging techniques that are now emerging to compete.

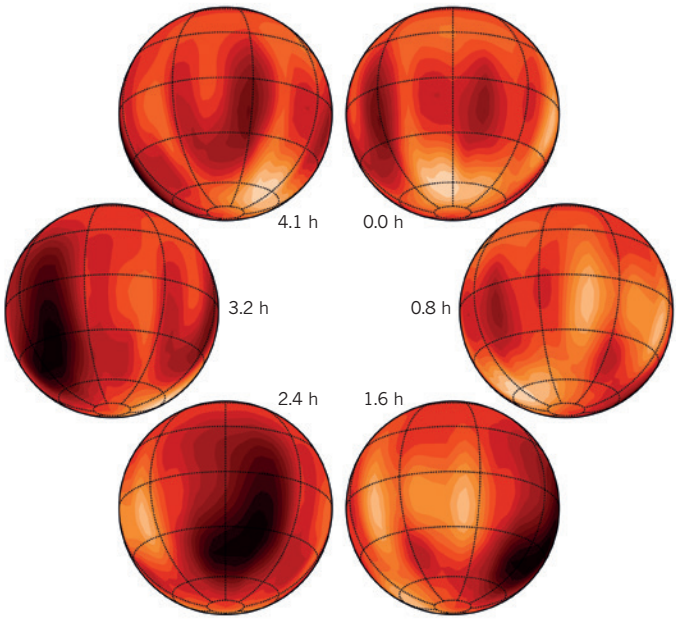


Figure 5 | Surface map of brown dwarf Luhman 16B. These maps are obtained by Doppler imaging and depict different epochs during the rotation of Luhman 16B. Large-scale cloud inhomogeneities are suggested by the dark patches at 2.4 hours and 3.2 hours. The rotation period of the brown dwarf is 4.9 hours. Reprinted with permission from ref. 76.

High-contrast imaging

Before the successful emergence of radial-velocity and transit methods, astronomers expected that high-contrast direct imaging that separates out the light of the planet and of the star, and provides photometric and spectroscopic data for each, would be the leading means of exoplanet discovery and characterization. A few wide-separation brown dwarfs and/or super-Jupiter planets were detected by this means, but the yield was meagre. The fundamental problem is twofold: the planets

are intrinsically dim, and it is difficult to separate out the light of the planet from under the glare of the star for planet–star separations like those of the Solar System. Imaging systems need to suppress the stellar light scattered in the optics that would otherwise swamp the planet's signature. The planet/star contrast ratio for Jupiter is ~10^{-9} in the optical and ~10^{-7} in the mid-infrared. For Earth, the corresponding numbers are ~10^{-10} and ~10^{-9}. These numbers are age, mass, orbital distance and star dependent, but demonstrate the challenge. Furthermore, contrast capabilities are functions of planet–star angular separation, restricting the orbital space that is accessible.

However, high-contrast imaging is finally emerging to complement other methods. It is most sensitive to wider-separation (~10–200 AU), younger giant exoplanets (and brown dwarfs), but technologies are coming online with which to detect older and less massive exoplanets down to around 1.0 AU separations for nearby stars (closer than ~10 parsecs)[10–13,77]. Super-Neptunes around M dwarfs might soon be within reach. Using direct imaging, Marois *et al.*[78,79] have detected four giant planets orbiting the A star HR 8799 (HR 8799b, HR 8799c, HR 8799d and HR 8799e) and Lagrange *et al.*[80] have detected a planet around the A star β Pictoris. The contrast ratios in the near infrared are about 10^{-4}, but capabilities near 10^{-5} have been achieved and performance near 10^{-7} is soon anticipated[10,11]. One of the results to emerge from the measurements of both the HR 8799 and β Pictoris planets is that to fit their photometry in the near-infrared from ~1.0 to 3.0 μm, thick clouds, even thicker than those seen in L-dwarf brown dwarf atmospheres, are necessary[81]. This re-emphasizes the theme that the study of hazes and clouds (nephelometry) has emerged as a core topic in exoplanet studies.

One of the most exciting recent measurements through direct imaging was by Konopacky *et al.*[82] of HR 8799c. Using the Ohio State Infrared Imager/Spectrometer (OSIRIS) on the 10-metre Keck II telescope, they obtained unambiguous detections between ~1.95 μm and 2.4 μm of both water and carbon monoxide in its ~1,000 K atmosphere. This $\lambda/\Delta\lambda = 4{,}000$ spectrum is one of the best obtained so far, but was enabled by the youth (around 30 million years), wide-angular separation and large mass (~5–10 M_J) of the planet.

Improvements in theory that are needed to support direct imaging campaigns mirror those needed for light curves, but are augmented to include planet-evolution modelling to account for age, metallicity or composition and mass variations. Most high-contrast instruments are focused on the near-infrared, so cloud physics and near-infrared line lists for likely atmospheric constituents will require further work. The reader will note again that most observations and measurements of exoplanet atmospheres have been for giants. There are a few for sub-Neptunes and super-Earths, but high-contrast measurements of Earths around G stars like the Sun are not likely in the near future[83,84]. The planet/star contrast ratios are just too low, although Earths around M stars might be within reach — if we get lucky. For now, giants and Neptunes are the focus, as astronomers hone their skills for an even more challenging future.

What we know about atmospheric compositions

The species we have, without ambiguity, discovered so far in exoplanet atmospheres are: water, carbon monoxide, sodium, potassium and hydrogen (H_2), with various ionized metals indicated in exoplanet winds. Expected, but as yet undetected, species include: ammonia, methane, nitrogen gas, carbon dioxide, hydrogen sulphide, phosphine, hydrogen cyanide, acetylene, ethylene, oxygen, ozone and nitrous oxide. The nature of the hazes and clouds inferred is as yet unknown. The atmospheres probed have temperatures from ~600 K to ~3,000 K. Good spectra are the essential requirements for unambiguous detection and identification of molecules in exoplanet atmospheres, and these have been rare. Determining abundances is also difficult, because to do so requires not only good spectra, but also reliable models. Errors in abundance retrievals of more than an order of magnitude are likely, and this fact has limited the discussion of abundances in this paper.

Nevertheless, with the construction of ground-based ELTs, the various campaigns of direct imaging[10–12], the launch of the JWST, the possible launch of the 2.4 m Wide-Field Infrared Survey Telescope (WFIRST)-AFTA[13], the various ongoing campaigns with Hubble and Spitzer, and with extant ground-based facilities, the near-term future of exoplanet atmospheric characterization promises to be even more exciting than its past. ■

Received 10 April; accepted 23 June 2014.

1. Mayor, M. & Queloz, D. A Jupiter-mass companion to a solar-type star. *Nature* **378,** 355–359 (1995).
This discovery paper inaugurated the modern era of exoplanet research.
2. Charbonneau, D., Brown, T., Latham, D. W. & Mayor, M. Detection of planetary transits across a Sun-like star. *Astrophys. J.* **529,** L45–L48 (2000).
3. Borucki, W. J. *et al.* Kepler planet-detection mission: introduction and first results. *Science* **327,** 977–980 (2010).
4. Moutou, C. *et al.* CoRoT: Harvest of the exoplanet program. *Icarus* **226,** 1625–1634 (2013).
5. Burrows, A. S. Spectra as windows into exoplanet atmospheres. *Proc. Natl Acad. Sci. USA* http://dx.doi.org/10.1073/pnas.1304208111 (2014)
This paper provides an unvarnished appraisal of current state-of-the-art exoplanet atmosphere characterization.
6. Brown, T. Transmission spectra as diagnostics of extrasolar giant planet atmospheres. *Astrophys. J.* **553,** 1006–1026 (2001).
7. Seager, S. & Sasselov, D. D. Theoretical transmission spectra during extrasolar giant planet transits. *Astrophys. J.* **537,** 916–921 (2000).
8. Fortney, J. J. *et al.* On the indirect detection of sodium in the atmosphere of the transiting planet HD209458b. *Astrophys. J.* **589,** 615–622 (2003).
9. Werner, M. W. *et al.* The Spitzer Space Telescope mission. *Astrophys. J.* **154** (Suppl.), 1–9 (2004).
10. Macintosh, B. *et al.* The Gemini Planet Imager: from science to design to construction. *Proc. SPIE* **7015,** 701518 (2008).
11. Beuzit, J.-L. *et al.* SPHERE: a planet finder instrument for the VLT. *Proc. SPIE* **7014,** 701418 (2008).
12. Suzuki, R. *et al.* Performance characterization of the HiCIAO instrument for the Subaru Telescope. *Proc. SPIE* **7735,** 773530 (2010).
13. Spergel, D. N. *et al.* Wide-field infrared survey telescope – astrophysics focused telescope assets WFIRST-AFTA final report. Preprint at http://arxiv.org/abs/1305.5422 (2013).
14. Madhusudhan, N., Knutson, H., Fortney, J. J. & Barman, T. Exoplanetary atmospheres. Preprint at http://arXiv.org/abs/1402.1169 (2014).
15. Burrows, A. & Orton, G. in *Exoplanets* (ed. Seager, S.) 419–440 (Univ. Arizona Press, 2010).
16. Fletcher, L. N. *et al.* Exploring the diversity of Jupiter-class planets. *Phil. Trans. R. Soc. A* **372,** 20130064 (2014).
17. Tinetti, G. Galactic planetary science. Preprint at http://arXiv.org/abs/1402.1085 (2014).
18. Tinetti, G. *et al.* in *Proc. Inter. Astronom. Union IAU Symposium 6* (S276) 359–370 (2011).
19. Guillot, T. On the radiative equilibrium of irradiated planetary atmospheres. *Astron. Astrophys.* **520,** A27–A39 (2010).
20. Burrows, A., Hubbard, W. B. & Lunine, J. I. The theory of brown dwarfs and extrasolar giant planets. *Rev. Mod. Phys.* **73,** 719–765 (2001).
21. Burrows, A. *et al.* A nongray theory of extrasolar giant planets and brown dwarfs. *Astrophys. J.* **491,** 856–875 (1997).
22. Deming, D. *et al.* Discovery and characterization of transiting super Earths using an all-sky transit survey and follow-up by the James Webb Space Telescope. *Publ. Astron. Soc. Pac.* **121,** 952–967 (2009).
23. Rothman, L. S. *et al.* The HITRAN 2008 molecular spectroscopic database. *J. Quant. Spectrosc. Radiat. Transf.* **110,** 533–572 (2009).
24. Tennyson, J. & Yurchenko, S. N. The status of spectroscopic data for the exoplanet characterisation missions. Preprint at http://arXiv.org/abs/1401.4212 (2014).
25. Hill, C., Yurchenko, S. N. & Tennyson, J. Temperature-dependent molecular absorption cross sections for exoplanets and other atmospheres. *Icarus* **226,** 1673–1677 (2013).
26. Freedman, R. S., Marley, M. S. & Lodders, K. Line and mean opacities for ultracool dwarfs and extrasolar planets. *Astrophys. J.* **174** (Suppl.), 504–513 (2008).
27. Sharp, C. M. & Burrows, A. Atomic and molecular opacities for brown dwarf and giant planet atmospheres. *Astrophys. J.* **168** (Suppl.), 140–166 (2007).
28. Lodders, K. Solar System abundances and condensation temperatures of the elements. *Astrophys. J.* **591,** 1220–1247 (2003).
29. Lodders, K. & Fegley, B. Atmospheric chemistry in giant planets, brown dwarfs, and low-mass dwarf stars. I. Carbon, nitrogen, and oxygen. *Icarus* **155,** 393–424 (2002).
30. Lodders, K. & Fegley, B. *The Planetary Scientist's Companion* (Oxford Univ. Press, 1998).
31. Burrows, A. & Sharp, C. Chemical equilibrium abundances in brown dwarf and extrasolar giant planet atmospheres. *Astrophys. J.* **512,** 843–863 (1999).
32. Schaefer, L. & Fegley, B. Chemistry of silicate atmospheres of evaporating super-Earths. *Astrophys. J.* **703,** L113–L117 (2009).
33. Lecavelier des Etangs, A., Pont, F., Vidal-Madjar, A. & Sing, D. Rayleigh scattering in the transit spectrum of HD 189733b. *Astron. Astrophys.* **481,** L83–L86 (2008).
34. Charbonneau, D., Brown, T., Noyes, R. W. & Gilliland, R. L. Detection of an

extrasolar planet atmosphere. *Astrophys. J.* **568,** 377–384 (2002).
This paper announced the first ever detection of a specific chemical species (sodium) in an exoplanet atmosphere.
35. Sing, D. K. *et al.* Gran Telescopio Canarias OSIRIS transiting exoplanet atmospheric survey: detection of potassium in XO-2b from narrowband spectrophotometry. *Astron. Astrophys.* **527,** 10 (2011).
36. Pont, F. *et al*. The prevalence of dust on the exoplanet HD 189733b from Hubble and Spitzer observations. *Mon. Not. R. Astron. Soc.* **432,** 2917–2944 (2013).
This paper is the most definitive study at primary transit pointing to the potential centrality of obscuring hazes in an exoplanet atmosphere.
37. Pont, F., Knutson, H. A., Gilliland, R. L. M. & Charbonneau D. Detection of atmospheric haze on an extrasolar planet: the 0.55–1.05 μ transmission spectrum of HD 189733b with the Hubble Space Telescope. *Mon. Not. R. Astron. Soc.* **385,** 109–118 (2008).
38. Burrows, A., Marley, M. M. & Sharp, C. M. The near-infrared and optical spectra of methane dwarfs and brown dwarfs. *Astrophys. J.* **531,** 438–446 (2000).
39. Grillmair, C. J. *et al*. Strong water absorption in the dayside emission spectrum of the planet HD189733b. *Nature* **456,** 767–769 (2008).
40. Howe, A. & Burrows, A. Theoretical transit spectra for GJ 1214b and other 'super-Earths'. *Astrophys. J.* **756,** 176–189 (2012).
41. Kreidberg, L. *et al*. Clouds in the atmosphere of the super-Earth exoplanet GJ1214b. *Nature* **505,** 69–72 (2014).
42. Deming, D. *et al*. Infrared transmission spectroscopy of the exoplanets HD 209458b and XO-1b using the Wide-Field Camera-3 on the Hubble Space Telescope. *Astrophys. J.* **774,** 95–112 (2013).
43. Marley, M. S., Ackerman, A. S., Cuzzi, J. N. & Kitzmann, D. in *Comparative Climatology of Terrestrial Planets* (eds Mackwell, S., Bullock, M. & Harder, J.) 367–391 (Univ. Arizona Press, 2013).
44. Showman, A. P. *et al.* Atmospheric circulation of hot Jupiters: coupled radiative-dynamical general circulation model simulations of HD 189733b and HD 209458b. *Astrophys. J.* **699,** 564–584 (2009).
45. Deming, D., Seager, S., Richardson, L. J. & Harrington, J. Infrared radiation from an extrasolar planet. *Nature* **434,** 740–743 (2005).
This paper was the first to demonstrate the potential of the Spitzer Space Telescope to measure the light of an exoplanet during secondary eclipse.
46. Charbonneau, D. *et al*. Detection of thermal emission from an extrasolar planet. *Astrophys. J.* **626,** 523–529 (2005).
47. Burrows, A., Hubeny, I. & Sudarsky, D. A theoretical interpretation of the measurements of the secondary eclipses of TrES-1 and HD 209458b. *Astrophys. J.* **625,** L135–L138 (2005).
48. Burrows, A. & Lunine, J. Astronomical questions of origins and survival. *Nature* **378,** 333 (1995).
49. Vidal-Madjar, A. *et al*. An extended upper atmosphere around the extrasolar planet HD209458b. *Nature* **422,** 143–146 (2003).
This paper was the first to discover signatures of winds emanating from exoplanets.
50. Lecavelier Des Etangs, A. *et al*. Evaporation of the planet HD 189733b observed in H I Lyman-á. *Astrophys. Astron.* **514,** 10 (2010)
51. Fossati, L. *et al*. Metals in the exosphere of the highly irradiated planet WASP-12b. *Astrophys. J.* **714,** L222–L227 (2010).
52. Kulow, J. R., France, K., Linsky, J. & Parke Loyd, R. O. Lyman-α transit spectroscopy and the neutral hydrogen tail of the hot Neptune GJ 436b. Preprint at http://arXiv.org/abs/1403.6834 (2014).
53. Ehrenreich, D. & Désert, J.-M. Mass-loss rates for transiting exoplanets. *Astron. Astrophys.* **529,** A136 (2011).
54. Linsky, J. L. *et al*. Observations of mass loss from the transiting exoplanet HD 209458b. *Astrophys. J.* **717,** 1291–1299 (2010).
55. Marley, M. S., Gelino, C., Stephens, D., Lunine, J. I. & Freedman, R. Reflected spectra and albedos of extrasolar giant planets. I. Clear and cloudy atmospheres. *Astrophys. J.* **513,** 879–893 (1999).
56. Sudarsky, D., Burrows, A. & Pinto, P. Albedo and reflection spectra of extrasolar giant planets. *Astrophys. J.* **538,** 885–903 (2000).
This paper provides a comprehensive theory of the reflection and albedo spectra of giant exoplanets as a function of orbital distance and planet mass.
57. Burrows, A., Sudarsky, D. & Hubeny, I. Spectra and diagnostics for the direct detection of wide-separation extrasolar giant planets. *Astrophys. J.* **609,** 407–416 (2004).
58. Barman, T. S., Hauschildt, P. H. & Allard, F. Phase-dependent properties of extrasolar planet atmospheres. *Astrophys. J.* **632,** 1132–1139 (2005).
59. Madhusudhan, N. & Burrows, A. Analytic models for albedos, phase curves, and polarization of reflected light from exoplanets. *Astrophys. J.* **747,** 25–40 (2012).
60. Esteves, L. J., De Mooij, E. J. W. & Jayawardhana, R. Optical phase curves of Kepler exoplanets. *Astrophys. J.* **772,** 51–64 (2013).
61. Rowe, J. *et al*. The very low albedo of an extrasolar planet: MOST space-based photometry of HD 209458. *Astrophys. J.* **689,** 1345–1353 (2008).
62. Burrows, A., Ibgui, L. & Hubeny, I. Optical albedo theory of strongly-irradiated giant planets: the case of HD 209458b. *Astrophys. J.* **682,** 1277–1282 (2008).
63. Knutson, H. A. *et al*. A map of the day-night contrast of the extrasolar planet HD 189733b. *Nature* **447,** 183–186 (2007).
Using one of the first exoplanet light curves, this paper determined a crude surface temperature map of an exoplanet.
64. Knutson, H. A. *et al*. 3.6 and 4.6 μ phase curves and evidence for non-equilibrium chemistry in the atmosphere of extrasolar planet HD 189733b. *Astrophys. J.* **754,** 22–37 (2012).
65. Majeau, C., Agol, E. & Cowan, N. B. A. Two-dimensional map of the extrasolar planet HD 189733b. *Astrophys. J.* **747,** L20–L24 (2012).
66. Knutson, H. A. *et al*. The 8 μ phase variation of the hot Saturn HD 149026b. *Astrophys. J.* **703,** 769–784 (2009).
67. Lewis, N. K. *et al*. Orbital phase variations of the eccentric giant planet HAT-P-2b. *Astrophys. J.* **766,** 95–117 (2013).
68. Cowan, N. B. *et al*. Thermal phase variations of WASP-12b: defying predictions. *Astrophys. J.* **747,** 82–98 (2012).
69. Crossfield, I. J. M. *et al*. A new 24 μ phase curve for υ Andromedae b. *Astrophys. J.* **723,** 1436–1446 (2010).
70. Seager, S., Whitney, B. A. & Sasselov, D. D. Photometric light curves and polarization of close-in extrasolar giant planets. *Astrophys. J.* **540,** 504–520 (2000).
71. Snellen, I. A. G., de Kok, R. J., de Mooij, E. J. W. & Albrecht, S. The orbital motion, absolute mass and high-altitude winds of exoplanet HD 209458b. *Nature* **465,** 1049–1051 (2010).
72. de Kok, R., *et al*. Detection of carbon monoxide in the high-resolution day-side spectrum of the exoplanet HD 189733b. *Astron. Astrophys.* **554,** A82 (2013).
73. Birkby, J. *et al*. Detection of water absorption in the day side atmosphere of HD 189733b using ground-based high-resolution spectroscopy at 3.2 μ. *Mon. Not. R. Astron. Soc.* **436,** L35–L39 (2013).
74. Brogi, M. *et al*. The signature of orbital motion from the dayside of the planet τ Boötes b. *Nature* **486,** 502–504 (2012).
75. Snellen, I. A. G. *et al*. The fast spin-rotation of the young extra-solar planet β Pictoris b. *Nature* **509,** 63–65 (2014).
76. Crossfield, I. J. M. *et al*. A global cloud map of the nearest known brown dwarf. *Nature* **505,** 654–656 (2014).
77. Burrows, A. A theoretical look at the direct detection of giant planets outside the Solar System. *Nature* **433,** 261–268 (2005).
78. Marois, C. *et al*. Direct imaging of multiple planets orbiting the star HR 8799. *Science* **322,** 1348–1352 (2008).
This paper represents the coming of age of the direct, high-contrast imaging technique of exoplanet discovery and characterization.
79. Marois, C., Zuckerman, B., Konopacky, Q. M., Macintosh, B. & Barman, T. Images of a fourth planet orbiting HR 8799. *Nature* **468,** 1080–1083 (2010).
80. Lagrange, A. M. *et al*. A probable giant planet imaged in the β Pictoris disk. VLT/NaCo deep L'-band imaging. *Astron. Astrophys.* **493,** L21–L25 (2009).
81. Madhusudhan, N., Burrows, A. & Currie, T. Model atmospheres for massive gas giants with thick clouds: application to the HR 8799 planets. *Astrophys. J.* **737,** 34–48 (2011).
82. Konopacky, Q. M., Barman, T. S., Macintosh, B. A. & Marois, C. Detection of carbon monoxide and water absorption lines in an exoplanet atmosphere. *Science* **339,** 1398–1401 (2013).
83. Kaltenegger, L., Traub, W. A. & Jucks, K. W. Spectral evolution of an Earth-like planet. *Astrophys. J.* **658,** 598–616 (2007).
84. Ehrenreich, D., Tinetti, G., Lecavelier Des Etangs, A., Vidal-Madjar, A. & Selsis, F. The transmission spectrum of Earth-size transiting planets. *Astron. Astrophys.* **448,** 379–393 (2006).

Acknowledgements The author acknowledges support in part under Hubble Space Telescope grants HST-GO-12181.04-A, HST-GO-12314.03-A, HST-GO-12473.06-A, and HST-GO-12550.02 and Jet Propulsion Laboratory/Spitzer Agreements 1417122, 1348668, 1371432, 1377197 and 1439064.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing financial interests. Readers are welcome to comment on the online version of this paper at go.nature.com/2e7lng. Correspondence should be addressed to A.S.B. (burrows@astro.princeton.edu).

REVIEW

doi:10.1038/nature13783

The role of space telescopes in the characterization of transiting exoplanets

Artie P. Hatzes[1]

Characterization studies now have a dominant role in the field of exoplanets. Such studies include the measurement of an exoplanet's bulk density, its brightness temperature and the chemical composition of its atmosphere. The use of space telescopes has played a key part in the characterization of transiting exoplanets. These facilities offer astronomers data of exquisite precision and temporal sampling as well as access to wavelength regions of the electromagnetic spectrum that are inaccessible from the ground. Space missions such as the Hubble Space Telescope, Microvariability and Oscillations of Stars (MOST), Spitzer Space Telescope, Convection, Rotation and Planetary Transits (CoRoT), and Kepler have rapidly advanced our knowledge of the physical properties of exoplanets and have blazed a trail for a series of future space missions that will help us to understand the observed diversity of exoplanets.

Over the past decade the study of exoplanets has rapidly evolved from a field that focused primarily on discoveries to one in which their characterization has taken central stage. These characterization studies include measurements of the mass, radius and bulk density of the planet, as well as the albedo, brightness temperature and composition of the atmosphere. The drivers of such studies are transiting exoplanets — planets whose orbital inclination with the line-of-sight is such that the planet periodically crosses in front of the star, resulting in a slight dip in its brightness.

The transit yields important characteristics about the exoplanet. From the dip in the stellar brightness during the transit, astronomers can derive the radius of the planet. When combined with the mass determination from spectroscopic Doppler measurements, this can be used to obtain the planet's mean density. The density gives us the first hints about the internal structure of the planet.

It is also possible to characterize the atmosphere of a transiting exoplanet using measurements taken at different points in its orbit. These measurements include in-transit spectra, phase variations of the exoplanet as it orbits the star and exoplanet occultations.

During the transit, a fraction of the starlight is absorbed by the planetary atmosphere. The planet's atmosphere thus imprints itself on the stellar spectrum giving us a glimpse of its composition. These kinds of measurements are called in-transit spectroscopy.

As the planet orbits, the illuminated portion of the sphere as seen from Earth changes, resulting in light variations — called a brightness phase curve — during the orbit. Measuring the phase curve enables astronomers to map the brightness distribution on the planet. High-precision photometric measurements can also detect other subtle variations in the light curve of the star. One effect is the so-called ellipsoidal variation that results from the distortion of the star due to tides raised by the planet. Another is the relativistic beaming effect, resulting from the star's reflex motion about the centre of mass of the system. Both of these phenomena provide independent measurements of the planet mass.

Finally, when the planet disappears behind the star, its reflected and radiated light is blocked. Phase and occultation measurements give us information about the brightness temperature, the albedo (the fraction of the incoming light that is reflected) and the exoplanet's spectral features.

Space-based instruments offer us the best means of characterizing transiting exoplanets. The superb temporal sampling (nearly continuous observations from months to years) allows us to extend the parameter space of transiting exoplanets to those with much longer orbital periods than can be discovered with ground-based telescopes. The exquisite photometric precision also allows astronomers to detect smaller planets than they could from the ground — exoplanets whose radii are comparable with that of the Earth, or even smaller. These high-precision measurements from space also make it easier to detect the minute signatures (less than 10^{-4}) of the reflected and radiated light from the exoplanet, as well as its atmospheric features. Finally, space instruments can access important wavelength regions, such as the infrared, that are simply not accessible from the ground.

Several space missions have contributed considerably to the discovery, study and characterization of transiting exoplanets. These include the Convection, Rotation and Planetary Transits (CoRoT) satellite, Kepler, the Microvariability and Oscillations of Stars (MOST) space telescope, the Hubble Space Telescope and the Spitzer Space Telescope. (See Review by Lissauer on page 336 for the contributions from the Kepler mission.)

CoRoT mission

CoRoT was the first spacecraft devoted to the discovery of transiting planets as well as the asteroseismic study of stars[1]. This mission was led by the French space agency Centre National d'Etudes Spatiales (CNES) along with contributions from the European Space Agency (ESA), Austria, Belgium, Brazil, Germany and Spain. Launched in December 2006, the 27-cm CoRoT telescope set the stage for NASA's larger and more ambitious Kepler mission. Over its lifetime, CoRoT observed more than 20 different star fields for up to a maximum of 150 days, collecting photometric data on approximately 175,000 stars.

CoRoT light curves are still being analysed for the signals of planets, but to date the mission has discovered 32 confirmed exoplanets, mostly gaseous giant planets (Fig. 1). For over two orders of magnitude in planet mass ($m \approx 0.3 - 20\ M_J$) these planets follow a tight mass–density power law:

$$\rho(\mathrm{gm\ cm^{-3}}) = (0.73 \pm 0.1) M_J^{(1.17 \pm 0.11)}$$

where M_J is the mass of Jupiter. This nearly linear relationship of density with mass reflects the fact that all objects with masses ranging from giant planets (~1 M_J) to low-mass stars (~100 M_J) have about the

[1]Thüringer Landessternwarte Tautenburg, Sternwarte 5, D-07778, Germany.

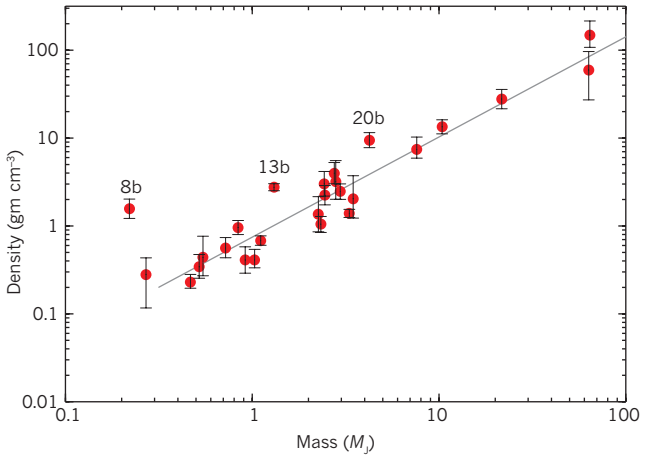


Figure 1 | Mass–density diagram for the CoRoT giant planets. Of the 32 exoplanets that CoRoT has discovered 26 are giant planets; the outliers CoRoT-8b, CoRoT-13b and CoRoT-20b are highlighted. The line represents a power law of the form $\rho(\mathrm{gm\ cm^{-3}}) = (0.73 \pm 0.1)\ M_J^{(1.17\pm0.11)}$, where M_J is the mass of Jupiter.

same radius. Thus, an increase in mass is accompanied by a proportional increase in density. This is expected as the pressure support in giant planets is provided by the electron-degeneracy pressure. Stars, however, have their pressure support provided by hydrogen burning under hydrostatic equilibrium.

There are two outliers that deviate from this power law relationship by almost 4σ: CoRoT-13b and CoRoT-20b. (The outlier at low planet mass CoRoT-8b is outside the mass range of the computed fit.) These planets have anomalous densities for their respective sizes. One hypothesis is that these may be the merger of two more 'normal' giant planets[2].

The CoRoT mission has contributed to exoplanet studies in a number of key areas.

Transiting planets around active stars

One of CoRoT's first exoplanet discoveries was CoRoT-2b[3]. This planet orbits a very active star, the surface of which is covered with cool spots. These spots produce photometric variability with an amplitude of about 2% that can mask the photometric dip caused by the transit events. CoRoT-2 is the most active star known to host a transiting exoplanet. Such a discovery would have been very difficult with ground-based measurements, and immediately demonstrated the value of those made from space. Planets around active stars are important for studying the interaction between the magnetic fields of the star and planet[4].

Cool giant planets

Before the launch of Kepler, CoRoT's planets accounted for half of transiting systems with orbital periods greater than 8 days. Noteworthy is CoRoT-9b — a planet in a 95-day orbit and the first transiting giant planet discovered with a moderate temperature[5]. Although most hot Jupiters have anomalously large radii that are typically 30–50% larger than that of Jupiter (R_J), CoRoT-9b has a quite normal radius (1.01 R_J). This supports the hypothesis that the anomalous radii of close-in giant planets are somehow associated with the high radiation flux coming from the host star[6]. CoRoT-9, with a brightness temperature of about 300 kelvin (K), has a 'balmier' temperature compared with other close-in transiting giant planets.

Oases in the brown-dwarf desert

Brown dwarfs are objects with masses (~13–60 M_J) that lie between giant planets and stars that are fusing hydrogen in their cores. Like planets, brown dwarfs undergo no core nuclear burning, but they do go through a short phase of deuterium burning. Ground-based Doppler measurements of stars have established that there is a paucity of such objects in relatively short-period orbits, the 'brown-dwarf desert'. CoRoT discovered at least three objects in this desert. In particular, CoRoT-3b with a mass of 22 M_J and the same radius as Jupiter[7] was the first brown dwarf to be characterized in terms of its mass and radius. This object may help our understanding of the relationship between giant planets and brown dwarfs and their respective formation scenarios.

A rocky super-Earth

The jewel in the crown of the CoRoT discoveries was the detection and characterization of the first transiting rocky planet, CoRoT-7b. The host star shows transit events every 20.5 hours with a photometric depth of a mere 0.03% caused by a planet with a radius 1.6 that of Earth (R_E)[8]. Ground-based spectroscopic Doppler measurements determined the mass of CoRoT-7b as 7.42 ± 1.21 M_E (where M_E is the mass of Earth)[9]. This corresponds to a density, ρ, of 10.4 ± 1.8 g cm^{-3}, possibly making the structure of CoRoT-7b closer to that of Mercury than that of Earth. CoRoT-7b gave us the first indication that rocky planets can have ultra-short-period orbits (orbital periods of less than 1 day). Shortly after its launch, Kepler discovered a virtual twin to CoRoT-7b — Kepler-10b[10]. This exoplanet has nearly the same orbital period (0.84 days) as CoRoT-7b, is slightly smaller (1.4 R_E), less massive (4.6 M_E), but has comparable bulk density (8.8 g cm^{-3}). Currently, the short-period rocky planet record holder is Kepler-78b, an Earth-sized planet that races around its host star in a mere 8.5 hours[11]. Recent ground-based Doppler measurements were able to confirm this as a rocky planet[12,13] that has an Earth-like density of 5.5 gm cm^{-3}.

Although CoRoT and Kepler were both devoted to the detection of planets by the transit method, the two missions were complementary in several ways. First, the target selection of Kepler was geared predominantly to stars like our Sun, whereas CoRoT obtained data on most stars in its field of view. As a result, CoRoT targets included a significant fraction of stars that were not Sun-like. About 35% of the CoRoT exoplanet discoveries orbit F-type stars, which are hotter and more massive than the Sun. By contrast, there is a 20% occurrence rate for planets around such stars found by Kepler. This may reflect the fact that CoRoT included more F-type stars as targets.

Second, the two missions searched for exoplanets in different directions in our Galaxy. Kepler looked at a field near the constellation of Cygnus, just off the plane of the Milky Way. Whereas, CoRoT looked at two regions in the sky (the 'eyes of CoRoT') in the galactic plane, one towards the centre of our galaxy and another towards the anti-centre. The galactic plane contains a larger number of young stars. Consequently, several of CoRoT's discoveries have been found around relatively young host stars. One of its most recent discoveries, CoRoT-32b (D. Gandolfi *et al.*, manuscript in preparation), which is about 25 million years old, might be one of the youngest discovered giant planets. Comparing the results of CoRoT and Kepler might help us to understand the role of the galactic environment in planet formation.

MOST space telescope

MOST is Canada's first space telescope and it was funded by the Canadian Space Agency, with ground-based and scientific support from Austria. It is a 15-cm-diameter telescope (often called 'The Humble Space Telescope') that was launched in 2003. Although designed to study stellar oscillations, it also delivered high-quality light curves on transiting exoplanets. MOST was the first space telescope to attempt to measure the optical albedo of an exoplanet and placed a low upper limit (less than 0.13) on the albedo of the transiting planet HD 209458b[14]. It also detected the transit of 55 Cancri e[15], a planet with nearly twice the diameter, 8.3 M_E, and with an ultra-short orbital period of only 0.74 days. The exoplanet 55 Cnc e was first discovered by ground-based Doppler surveys, but, with a photometric depth of only 0.04%, the transit would have been extremely difficult to detect with ground-based telescopes. Measurements from space were needed to show that this was an important transiting exoplanet. Ten years after its launch MOST continues to produce important results on transiting exoplanets.

It is worth mentioning that MOST, CoRoT and Kepler exemplify a strong spirit of international collaboration that is not often found in a highly competitive field such as exoplanet research. MOST data were provided to the CoRoT team to test on-board photometric reduction methods. Likewise, the CoRoT mission provided the Kepler team with early access to light curves so that they might better understand the influence of stellar activity. Each mission played an important part in the success of the subsequent one.

Hubble Space Telescope

Hubble has been a pioneering telescope for the study of exoplanetary atmospheres, with its suite of instruments such as the Space Telescope Imaging Spectrograph (STIS), the Near Infrared Camera and Multi-Object Spectrometer (NICMOS), the Advanced Camera for Surveys (ACS) and the Cosmic Origins Spectrograph (COS). Hubble was the first space-based facility to be employed in the study of exoplanets. It obtained, using STIS, what is arguably the finest light curve of an exoplanet transit, in this case the first known transiting planet HD 209458b[16]. This transit measurement established that space-based photometry could achieve the photometric precision needed to detect Earth-sized planets. It is possible that CoRoT and Kepler had an easier path towards approval owing to the pioneering measurements of Hubble. This telescope was also the first to detect an atomic species in an exoplanet atmosphere, namely sodium in HD 209458b[17]. Subsequently, in-transit spectroscopy with Hubble also detected the atomic species of hydrogen[18] and magnesium[19] in this exoplanet. Molecular oxygen may have also been found in HD 189733b[20] — another bright, nearby transiting exoplanet — using Hubble archive data.

We expected the atmospheres of giant planets to have molecules such as carbon monoxide, carbon dioxide, water and methane, much like the giant planets in our solar system. These features are best detected at infrared wavelengths. There have been reports of the detection of these molecules in the atmosphere of HD 209458b[21] and HD 189733b[22] using Hubble.

The detection of these molecules in the atmospheres of exoplanets, however, is controversial as the results might depend on the subtle details of how the data are reduced. For example, the resulting in-transit (transmission) spectra for several transiting exoplanets can be markedly altered when using different instrumental models to account for systematic errors[23]. Thus, the robustness of the detection of water, carbon dioxide and methane comes into question. This only underlines the difficulty, even when using instruments in space, of detecting molecular species in the atmospheres of exoplanets. The exoplanet field is truly 'pushing the envelope' when it comes to the detection of weak signals in data. We not only have to understand our instruments better, but also the star itself. For example, the presence of stellar spots and a poor knowledge of the physics of the stellar atmosphere will also affect our ability to detect atmospheric features of exoplanets[24]. The healthy discussion in the literature is a good example of how science is done: re-analyses of data and independent measurements to confirm sensational detections.

Hubble has also revealed that exoplanet atmospheres show evidence of Rayleigh scattering due to atmospheric haze. The transmission spectra of HD 189733b in the spectral range 0.550–1.050 μm obtained with the ACS of Hubble showed no indication of the expected sodium or potassium features, but rather a featureless slope that was consistent with Rayleigh scattering from clouds or haze in the upper atmosphere of the planet[25].

One of the more surprising results in the characterization of exoplanet atmospheres is that featureless atmospheric spectra may be common around Neptune-mass planets. The transmission spectrum of the transiting super-Earth GJ 1214b in the wavelength range 1–1.7 μm taken with the Wide Field Camera-3 of Hubble[26] is flat and lacks any obvious spectral features. The observed spectrum rules out a hydrogen-rich atmosphere and is consistent with an atmosphere abundant in water vapour, or with a considerable amount of clouds. A similar result was also obtained for the transiting GJ 436b Neptune-mass planet[27]. The detection of atmospheric absorption features in Neptune-like and super-Earth planets will be challenging and may require spectral observations taken at higher spectral resolution.

Spitzer Space Telescope

Spitzer is an infrared facility that was launched in 2009. It was the last of NASA's Great Observatories programme. This 0.85-m-diameter telescope operates at the long wavelengths of 3 μm to 180 μm. Planets orbiting close-in to their host stars have high equilibrium temperatures and thus most of their radiated light is at infrared wavelengths. For exoplanet studies, Spitzer has a marked advantage over Hubble, which operates mostly at optical and near infrared wavelengths.

Measurements were made of the transiting planet HD 189733b[28] using the 8-μm channel of the Infrared Array Camera (IRAC) on board Spitzer. These have produced a textbook example of the primary transit, the planet occultation and the variations due to the changing phases of the planet as it orbits the star (Fig. 2). The phase curve was used to map the temperature distribution on the exoplanet surface. This yielded a day and night temperature of 1,212 K and 973 K, respectively. Surprisingly, the map shows that the peak temperature for the planet is shifted by 16° from the sub-stellar point, the point on the planet directly under the star. The planet rotation is tidally locked to the orbital period so that the planet shows the same face to the star. Thus, one expects the

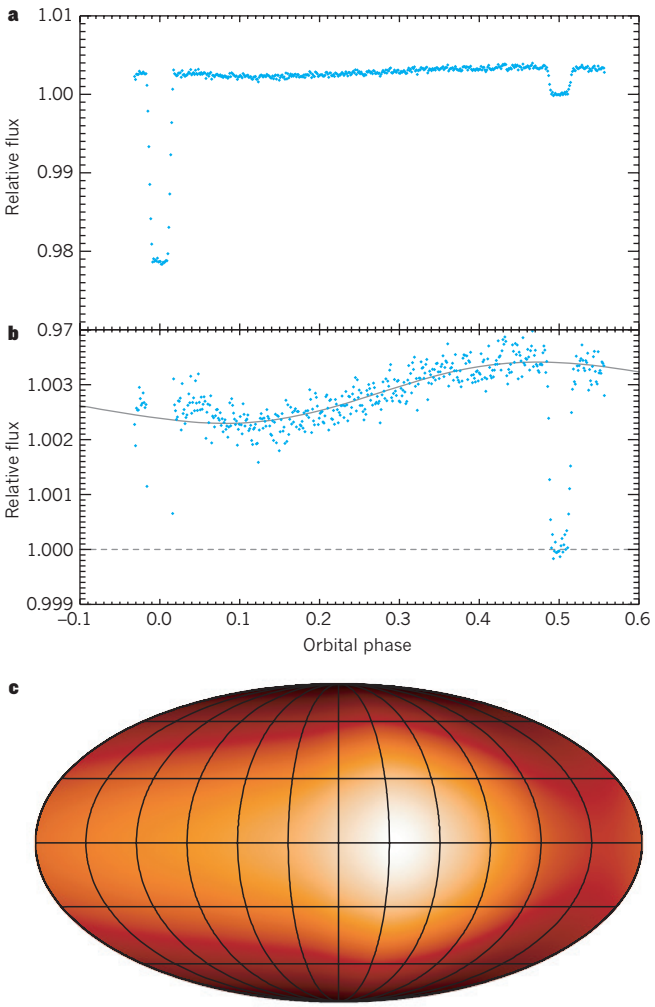


Figure 2 | Spitzer light curve for HD 189733. **a**, The observed phase variations with the transit and planet occultation visible. **b**, An expanded view to better show the phase variations. **c**, Brightness distribution of the surface of HD 189733b derived from the light curve. Figure reprinted with permission from ref. 28.

sub-stellar point to be the hottest, but it is not. One possible explanation for this phenomenon is that exoplanetary winds re-distribute the heat on the planet. This may be the first evidence of exoplanetary weather.

It is worth mentioning that exoplanet occultations and phase variations of an exoplanet at optical wavelengths were first detected by CoRoT[29]. These types of measurements made by Spitzer (infrared), Kepler (optical) and CoRoT (optical) have revealed that the albedo of most giant exoplanets is typically 10% or less, or comparable with that of the Moon. By comparison, Jupiter has an albedo of about 50%. Hot Jupiters are essentially dark and this is consistent with the early predictions from theoretical models of giant-planet atmospheres[30]. One exception to this is Kepler-7b, which has an almost Jupiter-like albedo of 0.32 (ref. 31).

In 2009, Spitzer exhausted its liquid helium cryogen, effectively stopping all observations in the long-wavelength channels. However, it continues to make important contributions using the 3.6-μm and 4.5-μm channels in the 'warm Spitzer' mode. For example, warm Spitzer independently detected the transit of 55 Cnc e[32] and determined its brightness temperature of 2,360 K[33]. So far, Spitzer has measured the temperatures of several dozen exoplanets.

Future space missions

Space-based investigations of exoplanets are currently in transition. Two important missions have experienced hardware failures. CoRoT lost its two remaining charge-coupled-device detectors in November 2012. Kepler lost its second reaction wheel in August 2013, thus compromising its ability to obtain the precise light curves needed for the detection of Earth-sized planets. Fortunately, the mission has been 'retooled' as Kepler 2, which will perform 'step and stare' observations on selected stellar fields in the ecliptic (see Review by Lissauer *et al.* on page 336).

Other space missions have fallen victim to budget constraints. The Canadian Space Agency has chosen to cut funding for MOST and it seems likely that 2014 will be its last year of operation. Spitzer has also been earmarked for shutdown by NASA, owing to budget woes. Both missions are still capable of making considerable contributions to exoplanet studies. The exoplanet community will certainly feel the impact of the loss of both these facilities as they made unique and unprecedented measurements that were just not possible from the ground. It is hoped that the respective agencies can find the monetary means to continue the operation of these important space missions.

Despite these setbacks, the future of space-based studies is bright, owing to a series of upcoming missions that will focus on relatively bright stars. The characterization of transiting planets is clearly best done on bright host stars. Doppler measurements for the accurate determination of the stellar mass requires high-resolution spectrographs, which can only be effectively used on relatively bright stars. Likewise, the detection of the reflected and radiated light from the planet and its atmospheric features require lots of photons to detect the minute signature of the atmosphere. Unfortunately, the transit planets discovered by CoRoT, Kepler and ground-based transit surveys are generally around faint stars, making characterization studies of transiting exoplanets challenging. Several approved space missions seek to remedy this situation.

The Bright Target Explorer (BRITE)-Constellation mission of Austria, Canada and Poland consists of six nanosatellites with telescopes of 3-cm aperture that will survey a few hundred of the brightest stars in the sky. BRITE Constellation may find transiting Neptune-sized planets around some of these stars. Nanosatellites are inexpensive and could find a cost-effective niche in exoplanet studies.

The Characterising Exoplanets Satellite (CHEOPS), a partnership between the ESA and the Swiss Space Agency, will be a 32-cm telescope that will search for transits of bright stars with exoplanets found by Doppler surveys. Its launch is planned for 2017.

NASA's Transiting Exoplanet Survey Satellite (TESS) will use four small telescopes to conduct a 2-year survey of nearby bright stars with V-magnitude brighter than 12th. (By comparison most CoRoT and Kepler targets have V-magnitude fainter than 12th.) Each field will be monitored for approximately 1 month before moving to the next field. The goal of TESS is to find targets for atmospheric studies using the James Webb Space Telescope (JWST). TESS is expected to be launched in 2017.

The Planetary Transits and Oscillations[34] (PLATO) mission of the ESA will also monitor several hundreds of thousands of the brightest stars using an array of 32 small telescopes. Its field of view will be about 20 times that of Kepler. Unlike TESS, PLATO will observe stars for several years so that terrestrial planets in the habitable zone can be discovered. Because the target stars are relatively bright, the Doppler measurement to determine the mass for these small planets will be significantly easier than it was for Kepler and CoRoT exoplanet candidates.

An important aspect of PLATO will be its asteroseismic component. Accuracy in planet mass and radius is limited by our uncertainty in the stellar parameters. For instance, transit measurements only yield the ratio of the planet to stellar radii. The precision and length of the PLATO light curves will allow astronomers to study the stellar oscillations of the host stars using asteroseismology. Analogous to how seismology is used to study Earth, asteroseismology can exploit stellar oscillations to determine accurate stellar and thus planetary parameters. With PLATO and ground-based follow-up measurements the planet mass will be determined with an error of 10%, the planet radius to within 2% and the age of the star (and thus planet) to within 10%. The planet–mass density measurements of PLATO should be able to distinguish whether a terrestrial planet is Mercury-like (with a large iron core), Earth-like (with an iron core and silicate mantle) or Moon-like (mostly silicates). It will also be able to determine accurate stellar ages for the host stars so that we can study the evolution of planetary systems as well. PLATO is scheduled for launch between 2022 and 2024.

Finally, JWST, successor to Hubble and Spitzer, is a 6.5-m telescope. It will be a significant improvement on Hubble and Spitzer for characterization studies of exoplanet atmospheres because of the larger aperture and the fact that it will be optimized for measurements in the infrared. JWST should be able to unambiguously detect the presence of water and methane in giant exoplanets as well as the atmospheric features of exoplanets as small as Neptune. This should establish whether the atmospheres of hot-Neptunes are truly featureless. It is scheduled for launch in late 2018 and has a nominal mission life of 5 years. JWST should be able to characterize the atmospheres of exoplanets discovered by TESS and possibly PLATO if JWST has an extended mission life.

There are a number of other space missions that have been proposed and are currently in the study phase. A description of all of these is beyond the scope of this short Review. With the approved and planned missions it can be said that we are entering a golden age of space-based studies of exoplanets that will ensure that an already vibrant and exciting field becomes even more so in the future. The scientific results from MOST, CoRoT, Kepler, Hubble and Spitzer set the foundation for the planning and development of these future missions. We anticipate a treasure trove of exciting exoplanet discoveries in the next decade that might help to answer the question of just how unique the planets in our Solar System are, and in particular Earth. ■

Received 13 May; accepted 15 July 2014.

1. Baglin, A. The CoRoT satellite in flight: description and performance. *Adv. Space Res.* **31,** 345–349 (2003).
2. Deleuil, M. *et al.* Transiting exoplanets from the CoRoT space mission. XX. CoRoT-20b: a very high density, high eccentricity transiting giant planet. *Astron. Astrophys.* **538,** A145 (2012).
3. Alonso, R. *et al.* Transiting exoplanets from the CoRoT space mission. II. CoRoT0-Exo-2b: a transiting planet around an active G star. *Astron. Astrophys.* **482,** L21 (2008).
4. Walker, G. A. H. *et al.* MOST detects variability on τ Bootis a possibly induced by its planetary companion. *Astron. Astrophys.* **482,** 691 (2008).
5. Deeg, H. J. *et al.* A transiting giant planet with a temperature between 250 K and 430 K. *Nature* **464,** 384–387 (2010).
6. Demory, B.-O. & Seager, S. Lack of inflated radii for Kepler giant planet candidates receiving modest stellar irradiation. *Astrophys. J.* **197** (Suppl.), 12 (2011).
7. Deleuil, M. *et al.* Transiting exoplanets from the CoRoT space mission. VI.

CoRoT-Exo-3b: the first secure inhabitant of the drown-dwarf desert. *Astron. Astrophys.* **491,** 889 (2008).
8. Leger, A. *et al*. Transiting exoplanets from the CoRoT space mission. VIII. CoRoT-7b: the first super-Earth with measured radius. *Astron. Astrophys.* **586,** 278 (2009).
This paper reports the discovery of the first transiting rocky super-Earth exoplanet.
9. Hatzes, A. P. *et al*. The mass of CoRoT-7b. *Astrophys. J.* **743,** 75 (2011).
10. Batalha, N. *et al*. Kepler's first rocky planet. *Astrophys. J.* **729,** 27 (2011).
11. Sanchis-Ojeda, R. *et al*. Transits and occultations of an Earth-sized planet in an 8.5-hour orbit. *Astrophys. J.* **774,** 54 (2013).
12. Howard, A. W. *et al.* A rocky composition for an Earth-sized exoplanet. *Nature* **503,** 381–384 (2013).
13. Pepe, F. *et al*. An Earth-sized planet with an Earth-like density. *Nature* **503,** 377–380 (2013).
Together with ref. 12, this paper reports the discovery of the first transiting rocky Earth-sized exoplanet.
14. Rowe, J. *et al*. The very low albedo of an extrasolar planet: MOST space-based photometry of HD 209458. *Astrophys. J.* **689,** 1345 (2008).
15. Winn, J. N. *et al*. A super-Earth transiting a naked-eye star. *Astrophys. J.* **737,** L18 (2011).
16. Brown, T. M., Charbonneau, D., Gilliland, R. L., Noyes, R. W. & Burrows, A. Hubble Space Telescope time series photometry of the transiting planet of HD 202458. *Astrophys. J.* **552,** 699 (2001).
17. Charbonneau, D., Brown, T. M., Noyes, R. W. & Gilliland, R. L. Detection of an extrasolar planetary atmosphere. *Astrophys. J.* **568,** 377 (2002).
18. Vidal-Madjar, A. *et al*. An extended upper atmosphere around the extrasolar planet HD 209458b. *Nature* **422,** 143–146 (2003).
19. Vidal-Madjar, A. *et al*. Magnesium in the atmosphere of the planet HD209458 b: observations of the thermosphere-exosphere transition region. *Astron. Astrophys.* **560,** A54 (2013).
20. Ben-Jaffel, L. & Ballester, G. E. Hubble Space Telescope detection of oxygen in the atmosphere of exoplanet HD 189733b. *Astron. Astrophys.* **553,** A52 (2013).
21. Swain. M.R. *et al*. Water, methane, and carbon dioxide present in the dayside spectrum of the exoplanet HD 209458b. *Astrophys. J.* **704,** 1616 (2009)
22. Swain, M. R., Vasisht, G. & Tinetti, G. The presence of methane in the atmosphere of an extrasolar planet. *Nature* **452,** 329–331 (2008).
23. Gibson, N. P., Pont, F. & Aigrain, S. A new look at NICMOS transmission spectroscopy of HD 189733, GJ-436 and XO-1: no conclusive evidence for molecular features. *Mon. Not. R. Astron. Soc.* **411,** 2199 (2011).
24. Csizmadia, S. Pasternacki, Th., Dreyer, C., Cabrera, J., Erikson, A., Rauer, H. The effect of stellar limb darkening values on the accuracy of the planet radii derived from photometric transit observations. *Astron. Astrophys.* **549,** A9 (2013).
25. Pont, F. *et al.* The prevalence of dust on the exoplanet HD 189733b from Hubble and Spitzer observations. *Mon. R. Astron. Soc.* **432,** 2917 (2013).
26. Kreidberg, L. *et al*. Clouds in the atmosphere of the super-Earth exoplanet GJ 1214b. *Nature* **505,** 69 (2014).
27. Knutson, H. A., Benneke, B., Deming, D. & Homeier, D. A featureless transmission spectrum for the Neptune-mass exoplanet GJ 436b. *Nature* **505,** 66–68 (2014).
28. Knutson, H. A. *et al*. A map of the day-night contrast of the extrasolar planet HD 189733b. *Nature* **447,** 183–186 (2007).
This paper documents the first brightness map of an exoplanet.
29. Snellen, I. A. G. de Mooij. Ernst, J.W. & Albrecht, S. The changing phases of extrasolar planet CoRoT-1b. *Nature* **459,** 543–545 (2009).
30. Sudarsky, D., Burrows, A. & Hunbeny, I. Theoretical spectra and atmospheres of extrasolar giant planets. *Astrophys. J.* **588,** 1121 (2003).
31. Demory, B.-O. *et al.* The high albedo of the Hot Jupiter Kepler-7b. *Astrophys. J.* **197,** 12 (2011).
32. Demory, B.-O. *et al*. Detection of a transit of the super-Earth 55 Cancri e with Warm Spitzer. *Astron. Astrophys.* **533,** A114 (2011).
33. Demory, B.-O. *et al*. Detection of a thermal emission from a super-Earth. *Astrophys. J.* **751,** L28 (2012).
34. Rauer, H. *et al*. The PLATO 2.0 Mission. *Exp. Astron.* (submitted); preprint at http://arxiv.org/abs/1310.0696 (2014).

Acknowledgements The author would like to warmly thank H. Rauer, J. Cabrera and S. Csizmadia for their valuable comments.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing financial interests. Readers are welcome to comment on the online version of this paper at go.nature.com/zohkcw. Correspondence should be addressed to A.P.H. (artie@tls-tautenburg.de).

REVIEW

doi:10.1038/nature13784

Instrumentation for the detection and characterization of exoplanets

Francesco Pepe[1], David Ehrenreich[1] & Michael R. Meyer[2]

In no other field of astrophysics has the impact of new instrumentation been as substantial as in the domain of exoplanets. Before 1995 our knowledge of exoplanets was mainly based on philosophical and theoretical considerations. The years that followed have been marked, instead, by surprising discoveries made possible by high-precision instruments. Over the past decade, the availability of new techniques has moved the focus of research from the detection to the characterization of exoplanets. Next-generation facilities will produce even more complementary data that will lead to a comprehensive view of exoplanet characteristics and, by comparison with theoretical models, to a better understanding of planet formation.

Astrometry is the most ancient technique of astronomy. It is, therefore, not surprising that the first (unconfirmed) detection of an extrasolar planet arose through this technique[1]. In 1984, another detection of a planetary-mass object around the nearby star VB 8 was reported, this time using speckle interferometry[2], but subsequent attempts to locate it were unsuccessful. It was finally Doppler velocimetry that delivered the first unambiguous detection of a very low-mass companion (HD 114762b[3]). However, because its minimum mass (11 M_J) is near the upper limit of the planetary mass range, the discoverers cautiously announced it was a brown dwarf. In 1992, a handful of bodies of terrestrial mass were found[4] and confirmed, by the measurement of timing variation, to orbit the pulsar PSR1257+12. Although very powerful, this technique was restricted to a small number of very particular hosts. Doppler velocimetry, instead, could be applied, with good results, to almost any 'quiet' star showing a reasonable amount of narrow absorption lines in its spectrum. The continuous improvement of this technique led, in 1995, to the discovery of the first giant planet around the Sun-like star 51 Pegasi[5] and marked the start of an intensive era of discoveries (see Review by Mayor *et al.* on page 328).

Since the discovery of 51 Peg b, microlensing, transit searches and direct imaging has delivered, together with Doppler velocimetry, an increasing number of planets and planetary candidates. Better instruments and improved detection limits have pushed our capabilities towards the detection of low-mass and small planets. Furthermore, the discovery of multi-planetary systems is the direct consequence of long-term, high-precision programmes. A new breakthrough was made thanks to the space-based transits searches Convection, Rotation and Planetary Transits (CoRoT)[6] and Kepler[7]. These missions have made a significant contribution to the statistical study of exoplanetary systems.

In this Review, we will discuss techniques and instruments that have contributed the most to our understanding of exoplanets. We will also provide an overview of present and future instrumentation, and describe how the field is moving from simple detection and statistical studies to the characterization of individual planets, their interior and their atmospheric composition.

Stellar radial velocities

Giant planets on short orbits induce radial-velocity variations in their host stars of several tens to a few hundreds of metres per second. Early Doppler velocimeters[8,9] delivered 200–500 m s^{-1} precision. With the use of a hydrogen-fluoride absorption cell the precision could be improved by one order of magnitude[10]. In the late 80s and early 90s an entire suite of new techniques and spectrographs[11–14] led to an improvement of the radial-velocity precision down to 3–15 m s^{-1}. This better precision led, in turn, to the discovery of 51 Peg b[5] and the era of giant-planet detection.

Would it be possible to detect terrestrial mass exoplanets by the Doppler technique? Some astronomers believed that improving the instrumental precision would be a key element[15]. Confirmation of this belief was provided by the discovery of μ Arae c in 2004 (ref. 16). At only 10 times the mass of Earth and with an orbit of 9.6 days, this planet produces a stellar radial-velocity pull of 3 m s^{-1} semi-amplitude. The detection of this tiny signal required a new generation of spectrographs, such as High Accuracy Radial Velocity Planet Searcher (HARPS)[17]. It represented the first step towards the detection and characterization of a vast population of Neptune-mass planets and super-Earths. The longer the temporal coverage and the better the instrumental precision, the smaller the radial-velocity signals (see for example the detection of α Cen B b[18]) that could be detected (Fig. 1).

The Doppler measurement consists of determining the wavelength of an identified spectral line and comparing it with the theoretical value it would have when transferred into the Solar System's rest frame. The Doppler equation links the measurement to the theoretical wavelength by the relative-velocity vector, finally delivering the projection of this vector in the direction of the line of sight (radial velocity). To increase the precision, the average, over several thousands of spectral lines, is computed. It should be noted, however, that the radial-velocity measurement is affected by several potential error sources that have been discussed extensively[10,14,19,20]. The main error sources are: photon noise[14,21]; instrumental errors[11,14,19]; spectrograph-illumination effects[22,23]; spectral contamination[19,24]; and stellar 'noise'[25–39], commonly referred to as stellar jitter. The term stellar jitter masks various stellar causes that produce radial-velocity effects at all timescales and of different magnitude. The discussion of all these effects lies beyond the scope of our Review. Nevertheless, it is important to be reminded that stellar jitter is probably the strongest limitation for Doppler velocimetry when aiming for sub-metre-per-second precision.

Present and future Doppler spectrographs need to address the mentioned limitations. As a first step, telescope size should be increased because high-spectral-resolution measurements are photon-starved, even for relatively bright targets. The gain obtained with a large telescope

[1]Observatoire Astronomique de l'Université de Genève, 51 Chemin des Maillettes, 1290 Versoix, Switzerland. [2]Swiss Federal Institute of Technology, Institute for Astronomy, Wolfgang-Pauli-Strasse 27, 8093 Zurich, Switzerland.

is, however, easily lost if spectral resolution is low. In fact, for unresolved spectral lines the measurement precision increases significantly with increasing spectral resolution[21]. In the photon-noise-limited regime the error ε_{bary} on the line-centre measurement can be estimated by:

$$\varepsilon_{bary} = \frac{\sigma^{1.5}}{\sqrt{2 \cdot I_0} \cdot EW} \cdot \sqrt{\left(1 - \frac{c}{2}\right)}$$

where σ is the measured width of the spectral line as seen through the spectrograph; $c = (I_{min} - I_0)/I_0$ is the measured line contrast; and $EW = \sigma c$ is the equivalent width. I_0 and I_{min} designate the photoelectron counts per resolution element in the continuum and the line minimum, respectively. It must be noted that the resolution element can be represented either by the detector pixel or by the wavelength unit as long as all the parameters are expressed in the same units. It is now commonly agreed that a spectral resolution of at least $R = \lambda/\Delta\lambda = 100{,}000$ should be used to guarantee the best precision on slowly rotating, quiet, solar-type stars. Spectral resolution and adequate line sampling not only allow us to achieve better signal-to-noise per spectral line, but also to reduce possible instrumental errors in both the radial-velocity measurement and the calibration process. To first order approximation, instrumental errors scale with the size of the resolution element (expressed in wavelength units). Unfortunately, with increasing telescope size, spectral resolution is a considerable driver of cost. For seeing-limited instruments the optical etendue E ($E = A \times \Omega$, the beam cross-section area times the solid angle) increases with the telescope size, and so does the instrument size if the spectral resolution is kept fixed[40]. In the era of 8-m class and extremely large telescopes (ELTs), this aspect has become a technical and managerial challenge that is nevertheless successfully addressed by employing novel optical design concepts[41–43].

All future projects for radial-velocity spectrographs (Table 1) aim to detect rocky planets in the habitable zone (the distance to the star at which liquid water can persist on the surface of the planet)[44] of a Sun-like and a low-mass star. To attain this objective they must be photon-efficient and precise to the sub-metre-per-second level. Photon efficiency is obtained with optimized designs and high-spectral resolution. High precision also requires the control of all instrumental effects. State-of-the-art instruments are therefore designed to be stable[17]. Gravity invariance and illumination stability of the spectrograph are crucial aspects that can only be obtained through a fibre feed[45–48]. Despite the intrinsic light-scrambling properties of optical fibres[49–51] it was soon realised that the illumination produced by a circular optical fibre depends on how the starlight is fed into the fibre. In other words, motions of the stellar image at the fibre entrance would produce a change in the illumination of the spectrograph and mimic a radial velocity effect. Considerable effort was invested in improving image scrambling by using double scramblers[13,49] and octagonal fibres[52,53]. Effective improvements have already been demonstrated on operational instruments[54,55].

Any instrumental effect that produces a distortion or a shift of the spectral line in the detector-pixel space will be interpreted, if not detected and recognized, as a wavelength change and thus a Doppler shift[20]. Two methods of tracking the instrumental profile changes have successfully been applied in the past. The first is to superimpose an absorption spectrum of a reference gas cell[10,14,56] on the stellar spectrum, such that the instrumental profile is continuously measured. This so-called self-calibration technique is particularly useful and effective in spectrographs with varying instrument profiles, as is the case for slit spectrographs. The disadvantages of this technique are the restricted bandwidth of the gas-cell spectrum, the loss of efficiency due to absorption in the light path, and the necessity for a sophisticated deconvolution process to recover the stellar spectrum and thus the radial velocity. This latter step requires the introduction of many additional parameters for spectral modelling. To obtain a given precision, higher signal-to-noise spectra must be acquired. The second method, the 'simultaneous reference technique'[13,17], is conceptually opposite. It assumes a stabilized instrumental profile that does not change between two wavelength calibrations of the spectrograph,

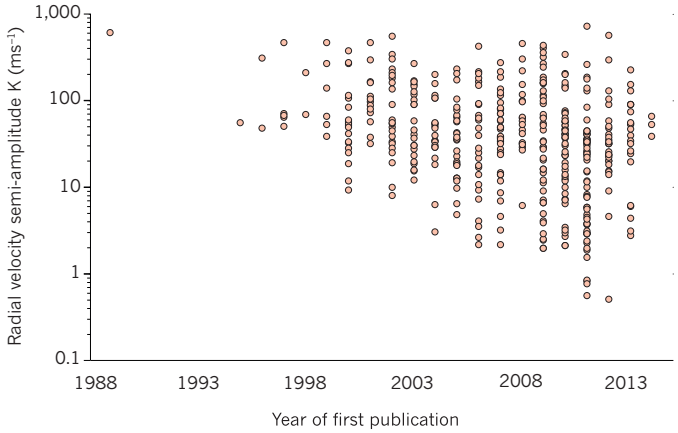


Figure 1 | Radial-velocity semi-amplitude of planetary-mass companions. All planets discovered by the Doppler technique from 1989 to present day are plotted. Remarkably, the detection limit improved by three orders of magnitudes in less than three decades. The underlying data were retrieved from http://www.exoplanets.org.

such that the determined relationship between the detector pixel and the wavelength remains valid over these timescales (typically a night). A second channel carrying a spectral reference is continuously fed to the spectrograph to monitor and correct for potential instrumental drifts or instrument profile changes. It must be guaranteed, however, that the changes that the scientific and the reference channels are subject to are identical over the timescale of one observing night. Therefore, the whole design of the instrument must be optimized for stability of the instrumental profile, requesting fibre feed and light scrambling, as well as pressure, mechanical, thermal and optical stability. The effort is compensated for by an unrestricted spectral bandwidth and the acquisition of an 'uncontaminated' scientific spectrum.

Although, in the case of the self-calibration technique, the instrument profile is supposed to be recoverable by deconvolution, there seems to be general agreement that low-order instrument-profile changes must, in any case, be avoided and that a stable instrument will eventually deliver more precise measurements. There is also agreement that better calibration sources are needed. The laser-frequency comb[57–62], when available at full potential, will provide the required calibration accuracy and precision. In the meantime, alternative sources are being developed, for example, passive Fabry–Pérot cavities[63–65] for simultaneous reference, or actively stabilized Fabry–Pérot systems for wavelength calibration[66].

The near-infrared wavelength region is becoming increasingly interesting for two other reasons. First, M dwarfs are much brighter in the infrared than in the visible[67]. These stars are cooler and thus their habitable zone lies closer to the host star. In addition the parent star is less massive. Potential habitable planets are, therefore, more easily detected by radial velocity[67]. Second, the influence of spots is strongly reduced in the near-infrared compared with the visible[68–70]. Furthermore, a comparison with radial-velocity determined in the visible wavelength range might help to discriminate a planet-induced velocity change from a stellar effect. For these reasons, many new instruments[71–76] (Table 1) will operate in the infrared wavelength domain. The use of adaptive optics[77] could be a means of reducing the size and cost of these instruments.

Transit photometry and spectroscopy

There are two approaches to detecting planetary transits: surveying as many stars as possible with one or several photometers in the hope of detecting new exoplanets through their transits, and photometrically following up planets discovered by Doppler velocimetry around their predicted inferior conjunction time. (The inferior conjunction denotes the orbital configuration where the planet lies between its host star and the observer; a transit occurs at the inferior conjunction if the orbital plane of the planet is aligned with the line of sight.) In the first method,

Table 1 | Non-exhaustive table of present (active) and future (approved) high-precision Doppler velocimeters

Instrument/technique	Telescope/observatory	Start of operations	Band (μm)	Spectral resolution	Efficiency (%)	Precision (m s^{-1})
Hamilton[180]/self-calibration	Shane 3 m/Lick	1986	0.34–1.1	30,000–60,000	3–6	3
UCLES[181]/self-calibration	3.9-m AAT/AAO	1988	0.47–0.88	~100,000	NA	3–6
HIRES[12]/self-calibration	Keck I/Mauna Kea	1993	0.3–1.0	25,000–85,000	6	1–2
CORALIE[13]/sim. reference	EULER/ESO La Silla	1998	0.38–0.69	60,000	5	3–6
UVES[182]/self-calibration	UT2–VLT/ESO Paranal	1999	0.3–1.1	30,000–110,000	4–15	2–2.5
HRS[183]/self-calibration	HET/McDonald	2000	0.42–1.1	15,000–120,000	6–9	3–6
HDS[184]/self-calibration	Subaru/Mauna Kea	2001	0.3–1.0	90,000–160,000	6–13	5–6
HARPS[18]/sim. reference	3.6 m/ESO La Silla	2003	0.38–0.69	115,000	6	< 0.8
FEROS-II[185]/sim. reference	2.2 m/ESO La Silla	2003	0.36–0.92	48,000	20	10–15
MIKE[186]/self-calibration	Magellan II/Las Campanas	2003	0.32–1.00	65,000–83,000 and 22,000–28,000	20–40	5
SOPHIE[187]/sim. reference	1.93 m/OHP	2006	0.38–0.69	39,000 and 75,000	4 and 8	2
CRIRES[188]/self-calibration	UT1–VLT/ESO Paranal	2007	0.95–5.2	~100,000	15	5
PFS[189]/self-calibration	Magellan II/Las Campanas	2010	0.39–0.67	38,000–190,000	10	1
PARAS[190]/sim. reference	1.2 m/Mt. Abu	2010	0.37–0.86	63,000	NA	3–5
CAFE[191]/sim. reference	2.2 m/Calar Alto	2011	0.39–0.95	~67,000	25	20
CHIRON[192]/self-calibration	1.5 m/CTIO	2011	0.41–87	80,000	15	<1
HARPS-N[54]/sim. reference	TNG/ORM	2012	0.38–0.69	115,000	8	<1
LEVY[193]/self-calibration	APF/Lick	2013	0.37–0.97	114,000–150,000	10–15	<1
EXPERT-III[194]/NA	2-m AST/Fairborn	2013	0.39–0.9*	100,000*	NA	NA
GIANO[71]/self-calibration	TNG/ORM	2014	0.95–2.5	50,000	20	NA
SALT–HRS[195]/self-calibration	SALT/SAAO	2014	0.38–0.89*	16,000–67,000*	10–15*	3–4*
FIRST[194]/NA	2-m AST/Fairborn	2014	0.8–1.8*	60,000–72,000*	NA	NA
IRD[73]/sim. reference	Subaru/Mauna Kea	2014	0.98–1.75*	70,000*	NA	1*
NRES/NA	6 × 1-m/LCOGT	2015	0.39–0.86*	53,000*	NA	3*
MINERVA/self-calibration	4 × 1-m/Mt. Hopkins	2015	0.39–0.86*	NA (Kiwispec)*	NA	1*
CARMENES[72]/sim. reference	Zeiss 3.5-m/Calar Alto	2015	0.55–1.7*	82,000*	10–13*	1*
PEPSI[196]/sim. reference	LBT/Mt. Graham	NA	0.38–0.91*	120,000–320,000*	10*	NA
HPF[74]/sim. reference	HET/McDonald	NA	0.98–1.40*	50,000*	4*	1–3*
CRIRES+/self-calibration	VLT/ESO Paranal	2017	0.95–5.2*	~100,000*	15*	<5*
ESPRESSO[42]/sim. reference	All UTs–VLT/ESO Paranal	2017	0.38–0.78*	60,000–200,000*	6–11*	0.1*
SPIROU[76]/sim. reference	CFHT/Mauna Kea	2017	0.98–2.35*	70,000*	10*	1*
G-CLEF[43]/sim. reference	GMT/Las Campanas	2019	0.35–0.95*	120,000*	20*	0.1*

For the spectral band and the spectral resolution the maximum value is given. The total efficiency has been extrapolated to include slit losses, and telescope and atmospheric throughput. The radial-velocity precision was estimated from published orbits or standard star's velocities. Historical instruments have not been listed. It is interesting to note that most of the planets discovered between 1995 and 2003 were detected using a small number of precision instruments: High Resolution Echelle Spectrometer (HIRES) at the 10-m Keck i telescope in Hawaii, CORALIE at the European Southern Observatory (ESO) 3.6-m telescope in La Silla, The Hamilton Spectrograph at the Shane 120-inch telescope at Lick, ELODIE at the 1.93-m telescope of the Haute-Provence Observatory[13], Advanced Fiber-Optic Echelle (AFOE) on the 1.5-m telescope at the Whipple Observatory[197], University College London Échelle Spectrograph (UCLES) at the Anglo-Australian Telescope (AAT), Coudé Echelle Spectrograph[198] on the 2.7-m telescope, the Sandiford Cassegrain Echelle spectrograph[198] on the 2.1-m telescope and the High-Resolution Spectrograph (HRS) at the Hobby-Eberly Telescope (HET), all of them at the McDonald Observatory. After 2003 the HARPS spectrograph opened a new window on the domain of super-Earths and mini-Neptunes by improving the radial-velocity precision below the metre-per-second level. Since then, the metre-per-second precision has become a 'standard' and a goal for most of the Doppler-velocimeter projects presented in the table. AAO, Australian Astronomical Observatory; APF, Automated Planet Finder; AST, Automatic Spectroscopic Telescope; CAFE, Calar Alto Fiber-fed Echelle; CARMENES, Calar Alto High-Resolution Search for M dwarfs with Exo-Earths with Near-Infrared and Optical Echelle Spectrographs; CFHT, Canada-France-Hawaii Telescope; CTIO, Cerro Tololo Inter-American Observatory; ESPRESSO, Echelle Spectrograph for Rocky Exoplanet and Stable Spectroscopic Observations; EXPERT-III, Extremely High Precision Extrasolar Planet Tracker III; FEROS-II, Fiberfed Extended Range Optical Spectrograph; FIRST, Florida Infrared Silicon Immersion Grating Spectrometer; G-CLEF, GMT-CfA Carnegie, Catolica, Chicago Large Earth Finder; GMT, Giant Magellan Telescope; HPF, Habitable-zone Planet Finder; HDS, High Dispersion Spectrograph; IRD, Infrared Doppler; LBT, Large Binocular Telescope; LCOGT, Las Cumbres Observatory Global Telescope Network; MINERVA, Miniature Exoplanet Radial Velocity Array; MIKE, Magellan Inamori Kyocera Echelle; NRES, Network of Robotic Echelle Spectrographs; OHP, Observatoire de Haute Provence; ORM, Observatorio del Roque de los Muchachos; PARAS, PRL Advanced Radial-velocity All-sky Search; PEPSI, Potsdam Echelle Polarimetric and Spectroscopic Instrument; PFS, Planet Finder Spectrograph; SAAO, South African Astronomical Observatory; SALT, Southern African Large Telescope; SOPHIE, Spectrographe pour l'Observation des Phénomènes des Intérieurs Stellaires et des Exoplanètes; SPIROU, SpectroPolarimètre Infra-Rouge; TNG, Telescopio Nazionale Galileo; UT, Unit Telescopes; UT2–VLT, Unit Telescope 2–Very Large Telescope; UVES, Ultraviolet and Visual Echelle Spectrograph. NA, not available or non-reliable information; sim. reference, simultaneous reference technique.
*Indicates design values.

the expected depth of the transit light curve dictates the photometric precision needed — for Jupiter-sized planets in transit across Sun-like stars the transits can be detected from the ground with amateur telescopes. Hot Jupiters, however, are only found orbiting about 1% of nearby solar-type stars[78], requiring observers to maximize the number of surveyed stars. Bright main sequence stars can be surveyed over a large fraction of the sky by wide-field cameras with small aperture telescopes and charge-coupled devices (CCDs), as illustrated by the Wide Angular Search for Planets[79] (WASP). Observations from a single location are limited, however, by the duration of the night. Time and sky coverage can be further improved with networks of small telescopes that relay from different longitudes, such as the Hungarian Automated Telescope Network[80] (HATNet) or the Trans-Atlantic Exoplanet Survey[81] (TrES).

The other strategy is to stare at crowded stellar fields. The 1.3-m telescope of the Optical Gravitational Lensing Experiment (OGLE) yielded the first discoveries of exoplanets through the transit method[82] by applying this strategy. The confirmation of these detections with velocimetry[83], however, required a large observational effort because of the faint optical

magnitudes (denoted V) of the stars surveyed (V = 14–16 mag). The first space missions dedicated to the search for transiting exoplanets, CoRoT[6,84] and Kepler[7], also stared at dense fields with high-cadence precise (relative) photometry (see the Reviews by Hatzes on page 353 and Lissauer *et al.* on page 336). Together, these satellites have surveyed several hundred thousand stars. Radial-velocity follow-up of CoRoT and Kepler exoplanet candidates remains difficult owing to the faint magnitudes of the host stars and the large number of targets needing follow-up. The faintness of the host stars also sets severe limits on the use of photon-starved techniques, such as transmission spectroscopy for the study of the planetary atmospheres. This technique requires bright host stars (Fig. 2), such as the hosts of planets discovered through velocimetry and later detected in transit. Only nine such exoplanets are known so far, but future space missions will search for more of these planets. In the meantime, and from the ground, planets transiting small stars such as M dwarfs are being looked for, because the transit signal is inversely proportional to the square of the stellar radius. The MEarth survey[85], composed of eight identical robotically controlled 40-cm telescopes with CCD detectors, found a super-Earth[86] that is especially amenable to follow-up atmospheric studies[87–91].

Studies of exoplanetary atmospheres

The hot gas giant HD 209458b was the first exoplanet captured in transit by two separate small telescopes[92,93], with a relative photometric precision of 0.2–0.4%. This transit was also the first exoplanet-related event observed from space: the 2.4-m Hubble Space Telescope measured the transit light curve to a precision of 110 p.p.m. per minute of observation[94]. The photometric observations of HD 209458b were obtained by integrating the stellar spectra collected before, during and after the transit by the Space Telescope Imaging Spectrograph (STIS)[95] CCD detector. These spectra were recorded with a medium-resolution (R = 5,540) grism of medium band pass, notably including the sodium doublet at 589 nm. The first transmission signature of an exoplanetary atmosphere was reconstructed from this data set by measuring, during the transit, an extra absorption of 200 p.p.m. in the sodium lines[96]. The far-ultraviolet channel of the STIS instrument, which collects ultraviolet photons with a multi-anode microchannel array (MAMA) detector, was used to observe the transit of HD 209458b over the stellar Lyman-α emission of atomic hydrogen at 121 nm. These measurements led to the discovery of an extended upper atmosphere to the planet[97].

HD 209458b remained, for quite some time, the only known transiting exoplanet. By the time additional transiting exoplanets were announced (in 2004), STIS had experienced a power-supply failure. The instrument was only repaired in 2009 during the last servicing mission of Hubble. Arguably, the main effect of the STIS failure was to shift the field of exoplanetary atmospheres into the infrared. After 2004, and despite successful attempts to record precise transit light curves with the Advanced Camera for Surveys on board Hubble[98], the 85-cm Spitzer Space Telescope became the prime observatory not only for transits, but also for eclipses of planets by their stars, which can occur at superior conjunctions (the orbital configuration opposite the inferior conjunction, when the planet passes behind the star). Broadband photometry of these eclipses with the Infrared Array Camera (IRAC)[99] on Spitzer revealed the thermal emission from exoplanets, the first example of direct detection of light from a planet orbiting a star[100–102]. The instrument has four broadband infrared channels collecting light on two detectors made of indium antimonide (3.6 μm and 4.5 μm channels) and arsenic-doped silicon (5.8 μm and 8.0 μm channels).

The first infrared observation of a planetary transit[103] was obtained with the Multiband Imaging Photometer for Spitzer (MIPS) at 24 μm. These observations were limited by the low stellar flux in the mid-infrared. Furthermore, transit observations in the near infrared exhibited large instrumental effects, precluding the detection of molecular signatures. Both photometry with IRAC[104–106] and spectroscopy with the Near-Infrared Camera and Multi-Object Spectrometer (NICMOS)[107] on Hubble yielded non-reproducible results or were of insufficient quality

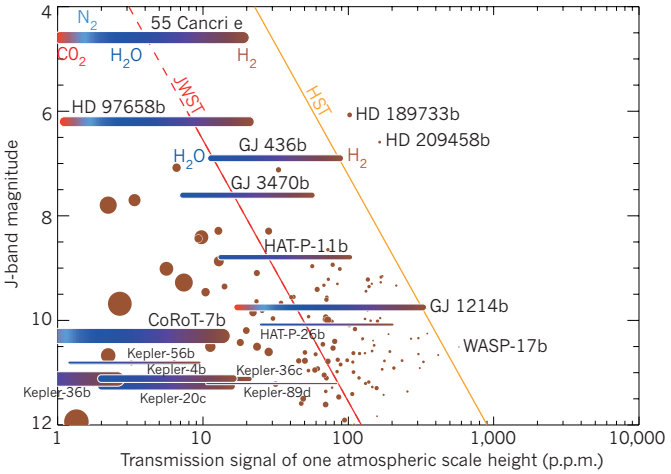


Figure 2 | Detectability of planetary atmospheres. The signal of one atmospheric scale height seen in transmission during transit is plotted against the stellar J magnitude. The signal is calculated in parts per million (p.p.m.) as $2 \times 10^6\ (\Delta F/F)(H/R_p)$, where $\Delta F/F$ is the transit depth and H is the atmospheric scale height. This quantity scales here with the equilibrium temperature of the planet and is inversely proportional to the acceleration of gravity at the surface of the planet and the mean molar mass (μ) of the atmosphere. The atmospheric signal is proportional to the planet mean density. The size of the circles (and the thickness of the colour bars) scales with the density to show this effect. For giant exoplanets (brown circles), the atmosphere is assumed to be primarily composed of molecular hydrogen (H_2) and helium ($\mu = 2.3\ g\ mol^{-1}$). For lower-mass planets, such as Neptunes ($10 < M_p < 60\ M_E$, where M_E is the mass of Earth and M_p is the mass of the planet) and super-Earths ($M_p < 10\ M_E$), the atmospheric composition is unknown and the colour bar extents represent all possible signal values assuming hydrogen and helium ($\mu = 2.3\ g\ mol^{-1}$, brown) and water (H_2O, $\mu = 18\ g\ mol^{-1}$, blue) dominated atmospheres for Neptunes, and molecular nitrogen (N_2, $\mu = 28\ g\ mol^{-1}$, light blue), and carbon dioxide (CO_2, $\mu = 44\ g\ mol^{-1}$, red) dominated atmospheres for super-Earths, in addition to the two earlier types. Approximate Hubble Space Telescope (HST) and JWST 3-σ detection limits (orange and red lines, respectively) are shown. Only super-Earths and Neptunes with a mass determined to better than 20% are represented.

for unambiguous interpretation[108–110]. Eclipse spectroscopy of the dayside emission of HD 189733b obtained with the third instrument on Spitzer, the Infrared Spectrograph (IRS)[111] providing low-resolution (R = 80) and spectral coverage from 5 μm to 14 μm, also had to be corrected for instrumental effects[112]. The IRS data nonetheless provided evidence for molecular absorption in an exoplanet atmosphere[113]. Unfortunately, the use of IRS was terminated after Spitzer ran out of cryogen in May 2009. Meanwhile, Spitzer continues observing with IRAC 3.6-μm and 4.5-μm channels, now commonly used to obtain precise transit light curves of exoplanets down to the super-Earth size regime[114,115].

Ground-based atmospheric characterization of exoplanets advanced through the use of high-resolution spectrographs. The signature of sodium in the atmosphere of HD 209458b was found[116] in data taken with the High Dispersion Spectrograph (R = 45,000) at the Subaru 8-m telescope[117]. The technique, differential spectroscopy, involves calibrating the signal in the spectroscopic features with the continuum signal in the vicinity of the features. The 'absolute' transit depth is lost, but the transmission signal can be retrieved, assuming that telluric absorption can be sufficiently calibrated. Another method is to calibrate the wavelength-dependent signal using other stars within the field of view of the instrument. This can be achieved in spectrophotometry for systems with nearby reference stars[118,119] or in spectroscopy with slit masks positioned on the target and on several reference stars in the field[87]. A breakthrough was made possible by the Cryogenic High-Resolution Infrared Echelle Spectrograph (CRIRES) on the Very Large Telescope (VLT). Its high resolution (R = 100,000), although over a narrow (50 nm) wavelength infrared region, allows tracking of the wavelength shift of individual spectral features composing molecular bands of water, carbon monoxide or carbon

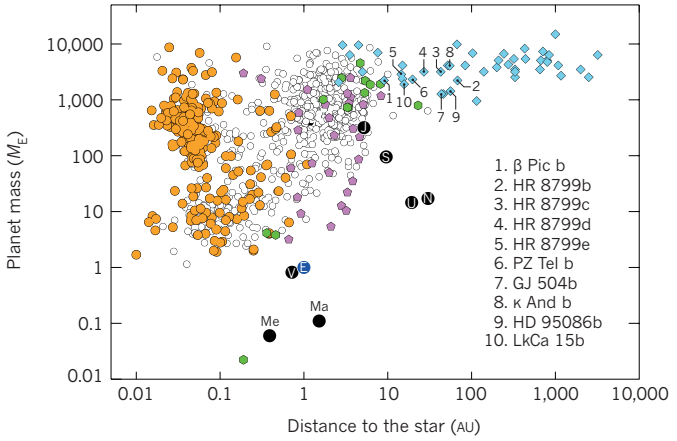


Figure 3 | Mass and semi-major axis of known planets. Planetary mass is plotted as a function of the semi-major axis (the distance to the host star). Solar-system planets are shown by black circles (Me, Mercury; V, Venus; Ma, Mars; J, Jupiter; S, Saturn; U, Uranus; N, Neptune) and Earth (E) is in blue. Exoplanets detected with different techniques and instrumentation are shown. Doppler velocimetry (white), transit with a measured mass (orange), direct imaging (blue), microlensing (pink), and pulsation timing (green). Among the direct-imaging planets, only ten (labelled) were found within 100 AU of their host and a mass ratio between the companion and its host star $q < 0.02$. Data underlying this plot were retrieved from the Exoplanet Encyclopaedia[199].

dioxide that are present in the atmosphere of the planet as the planet orbits the star[120–122]. The method works for transiting and non-transiting planets alike, giving access to the brightest exoplanetary systems, such as that of τ Boötis[123]. Its application to the directly imaged planet β Pictoris b[124] led to the determination of the spin velocity of the planet[125].

The refurbishment of Hubble in May 2009 enabled the recovery of STIS capabilities and the start of operations of both the Cosmic Origins Spectrograph[126] (COS) in the far-ultraviolet and the Wide-Field Camera-3 (WFC3)[127] in the near-infrared. COS and STIS provided observations and confirmation of the atmospheric mass loss from HD 209458b in the singly ionized carbon lines at 133 nm[128]. These measurements were extended to other exoplanets[129–133]. Visible STIS spectra revealed atomic signatures and the presence of light scattering processes in the upper atmospheric layers of HD 189733b[129,134], and observations of the eclipse of HD 189733b with a low-resolution grating from 290 nm to 570 nm also yielded the first chromatic measurements of a planetary albedo[135]. The WFC3 infrared channel was successfully used for slitless grism spectroscopy of exoplanetary transits in the near-infrared, achieving near-photon-noise transmission spectroscopy of super-Earths, Neptunes and gas giants between 1.1 μm and 1.7 μm[90,136–143], and detecting the 1.38-μm water band in some of these planets.

Direct imaging and astrometry

Despite many years of technological development, the search for ideal targets, improved analysis algorithms and investment in observing time on leading telescopes, it was not until 2008 that the first direct images of an exoplanetary system around a star were obtained. The multi-planet system HR 8799, with all planets[144] orbiting the intermediate mass host star in the same rotation sense, was a remarkable (and a lucky) breakthrough. Interpretation of the contemporaneous discovery of a faint point source around the debris disk host star Fomalhaut[145] has turned out to be more complex than anticipated[146]. Finally, at the end of 2008, a giant planet was found around β Pic[124,147–149] within the prototypical debris disk first imaged in the early 1980s[150]. These discoveries were preceded by several others (some of which were spurious), often around very young objects still in the process of becoming a star. For example, the companion to 2MASSWJ 1207334-393254 (a very young brown dwarf) was discovered[151] through adaptive-optics-assisted near-infrared imaging with the NACO instrument on the VLT. This discovery was notable because the system is very young, making detection of a self-luminous planetary mass object easier; the central host object is of very low mass and thus of modest luminosity relative to the planetary mass companion; and NACO is equipped with an infrared wave-front sensor, which is important to allow observations of this class of cool primaries. However, the mass ratio (q) of the brown dwarf to the companion is consistent with many examples of binary star systems of higher mass. So far, there have been 10 objects found within 100 AU of their host with a mass ratio between the companion and the host star of $q < 0.02$ (Fig. 3; http://exoplanet.eu/). These restrictions suggest that they may have formed like planets in our Solar System, but this is not at all certain. There are dozens of objects that have larger mass ratios (particularly around very low-mass primaries), as well as objects with low-mass ratios, but found at larger radii (out to more than 1,000 AU). One major caveat to these studies is that the masses are inferred from theoretical models[152] based on the shape of the spectral energy distribution and luminosity, as well as knowledge about the central star (primarily age, but also composition).

State-of-the-art instruments require advanced adaptive optics to correct for the blurring effects of Earth's atmosphere[153]. Although the diffraction limit improves at shorter wavelengths, high performance adaptive optics are more challenging, leading to compromises for instrument design between 0.5–5.0 μm. Even at the diffraction limit of an 8-m class telescope, it is only possible to reach orbital separations of 3 AU at 1.65-μm wavelength for stars out to a 50 pc distance. The younger a planet is, the hotter and brighter it is, making its detection and characterization easier. Nearby stars tend to be old (1–3 gigayear) and the youngest objects, which are more rare, are located at greater distances. Thus, another compromise needs to be found between available target sample and ease of detection, which translates directly into a balance between detectable mass (better for younger, more distant objects) and orbital separation (better for nearby stars). Results so far suggest that massive gas-giant planets ($> 2\ M_J$) are rare at large orbital radii[154] (for example, beyond 50 AU). However, new instruments utilizing extreme adaptive optics (resulting in an increase of hundreds to thousands of actuators controlling the shape of the deformable mirror, for example, the Spectro-Polarimetric High-Contrast Exoplanet Research (SPHERE)[155] instrument and the Gemini Planet Imager (GPI)[156]) will improve the inner working angle that can be reached at all wavelengths of operation, although in particular it will open up the possibility of Strehl ratios above 30% in the red visible[157]. It is also worth mentioning that great improvements in data acquisition modes and analysis software (differential imaging through angular, polarimetric and spectral difference) have greatly enhanced planet-detection capabilities[158–162]. In addition, the development of diffraction-suppression optics continues — as observations are contrast-limited close to the star. In the photon-noise limit, which is not often reached even around early type bright stars, sparse aperture masking[163] and coronography can improve the achievable contrast limit using techniques such as apodizing phase plates[164], vector vortex[165], phase-induced amplitude apodization[166] and classical Lyot coronography[167]. Marked improvements in diffraction suppression, stability and quality of adaptive optics, as well as in post-processing algorithms, are needed to reach the fundamental background-limited sensitivity close to the diffraction limit. The inner working angle, at which the background limit is reached, is 10 times larger than the diffraction limit. The implementation of low-noise infrared wave-front sensors is another key area of development, particularly in their application to imaging surveys of fainter lower-mass stars and brown dwarfs. Building the observational data to constrain the frequency of planets as a function of planet mass, orbital separation and primary-star mass will provide powerful tests for theories of planet formation.

The James Webb Space Telescope (JWST) will launch in 2018 and will provide powerful capabilities for direct imaging, including coronagraphy. All of its instruments will make great contributions to finding and characterizing exoplanets resolved from their host stars, including some of those already known today. In particular, its short-wavelength imager, Near Infrared Camera (NIRCam), will be able to detect planets

below the mass of Saturn beyond 30 AU around close-by stars. The Near-Infrared Imager and Slitless Spectrograph (NIRISS) will utilize a sparse aperture mask to detect bright companions below the diffraction limit at 1–2.3 μm wavelength. It will be particularly useful for surveys of very young stars for which planetary companions will be brightest relative to the central star. The Mid-Infrared Instrument (MIRI), the long-wavelength camera/spectrograph on JWST, will provide additional characterization of planetary atmospheres from 5 μm to 28 μm, and the Near-Infrared Spectrograph (NIRSpec, 1–5 μm) will be equipped with an integral field spectrograph that is capable of providing high-quality spectra of close companions.

Although JWST will be the most powerful telescope ever in terms of infrared sensitivity, it will not provide enhanced spatial resolution compared with the current generation of 6–10-m telescopes and will not provide unique capabilities for high-contrast imaging at inner working angles below 0.1 arcsec. Because we know that the distribution of giant gaseous planets rises with orbital radius out to 3 AU, and because massive gas giants are rare beyond 30 AU, it is likely that most Jupiter-mass planets will be found at intermediate separations. The next generation of ELTs will enable us to cross the 10 AU threshold in angular resolution of accessible targets, pushing the detectable separation down to 3 AU and enabling vast synergies between Doppler velocimetry and astrometry. The Large Binocular Telescope Interferometer (LBTI) is the first optical telescope with an effective resolution of a 22.8-m baseline[168], although it is not a filled aperture, thus limiting its sensitivity. The European ELT (E-ELT), with its aperture of 39 m, will integrate a suite of imaging and spectroscopic instruments (HARMONI, MICADO, METIS and eventually EPICS) to enable efficient imaging of exoplanets at diffraction-limited inner working angles below 0.1 arcsec. Similar instruments are planned for the two other ELT projects — the Thirty Meter Telescope (TMT) and the Giant Magellan Telescope (GMT). Considerable thought is being devoted to working out how to reach the fundamental background limit when approaching the diffraction limit, which, along with sensitivity and spatial resolution of these ELTs, would represent a major breakthrough. The removal of so-called quasi-static speckles is the key; this can be achieved, in principle, through sophisticated calibration schemes for adaptive-optics systems to enable the commanded removal of speckles, or equally by sophisticated analysis of the wavefront sensor camera data and telemetry to analyse residual errors in post-processing[169]. In predicting the performance of these future telescopes some take a conservative approach, whereas others believe they could reach the ultimate limit. Either way, these ELTs will represent a huge breakthrough in the capacity to directly image planets around nearby stars. If the technological challenges are mastered, the E-ELT will have a reasonable chance of obtaining a direct image of a super-Earth within 1 AU of the nearest stars[170].

In late 2013, the European Space Agency launched Gaia, which has the ability to reach micro-arcsecond astrometric precision. Owing to better performances, this mission will allow the exploration of a wider parameter space to detect motions in the plane of the sky due to the orbit of the host star and planet around a common centre of mass. As the precision will fall for fainter stars, Gaia will be sensitive to the lowest-mass planets only around stars in the solar neighbourhood, but will detect hundreds — if not thousands — of gas-giant planets within hundreds of parsecs[171]. Furthermore, this will open up the synergistic possibility to directly image some of these objects, providing ground-truth for models of their evolution. Ground-based astrometry will also play a part, as could JWST and other facilities. For instance, ground-based direct imaging could deliver astrometric measurements at 100 micro-arcsecond precision[172].

A bright and multi-technique future

In the past, our focus was on discovering new exoplanets and acquiring statistics about their diversity, which, in turn, concerned mainly external (orbital) parameters (Fig. 3). Now, interest is moving towards the detailed characterization of specific planets and planetary systems. Orbital parameters, host-star characteristics, synchronization and planetary spin, irradiation, planet density and internal structure, atmospheric composition, and physical conditions must be characterized in order for us to understand the formation processes and the observed diversity.

Increasing the number of targets amenable to further characterization is the prime goal of several dedicated space mission projects: the extension of the Kepler mission (K2)[173], NASA's Transiting Exoplanet Survey Satellite (TESS)[174] and ESA's Planetary Transits and Oscillations of Stars mission (PLATO)[175] will obtain photometric measurements of bright stars located almost everywhere in the sky, and thus find many new transiting planets around bright stars. These space missions will be complemented by new ground-based surveys dedicated to the search for transits across different types of stars, for example, the Next Generation Transit Survey (NGTS), the Search for Habitable Planets Eclipsing Ultra-cool Stars (SPECULOOS), the Exoplanets in Transit and Their Atmosphere (ExTrA) and the Multi-Site All-Sky Camera (MASCARA)[176].

The planets transiting bright stars will enable follow-up observations and characterization of the planets by other techniques. The Swiss-ESA spacecraft CHEOPS (Characterising Exoplanets Satellite)[177,178] will, by transit photometry, measure precise radii and bulk densities of known exoplanets and select the best-suited targets for atmospheric characterization by future spectrographs from space or on large ground-based telescopes. JWST will have unprecedented thermal infrared sensitivity. Its four instruments will, in addition to the direct imaging of planets, attempt transit observations at low-to-medium-resolution (R = 100–1,500) in the near- and mid-infrared domain for atmospheric characterization. Whereas several of the known hot gas giants will be amenable to detailed studies with JWST (Fig. 2), additional low-mass targets, Neptunes, super-Earths and Earth-like planets, will be delivered by TESS, CHEOPS and PLATO.

Atmospheric characterization of transiting and non-transiting exoplanets has already been initiated with current ground-based direct imaging and resolved spectroscopy as well as high-resolution spectrographs (for example, CRIRES and HARPS). These capabilities will be considerably extended with the upcoming generation of visible and near-infrared instruments equipping 4-m to 8-m telescopes. The advent of ELTs such as E-ELT, TMT and GMT, in combination with high spatial and spectral resolution, will amplify this tendency and open up a new parameter space, for instance by enabling the detection of bands of molecular oxygen on super-Earths transiting M dwarfs[179].

Techniques such as radial-velocity, photometry, astrometry, imaging and spectroscopy will all contribute to the field of exoplanets. Whereas in the past the groups using these techniques seemed to be in competition, now, in view of achieving a comprehensive understanding of the 'new worlds' we are looking for, the results they produce are highly complementary. The more mature the field becomes, the more we understand that we will not find another Earth with one single mission, but only with the combination of all the tools that are offered to us over the next decades. ■

Received 26 May; accepted 15 July 2014.

1. Van de Kamp, P. Parallax, proper motion, acceleration, and orbital motion of Barnard's star. *Astron. J.* **74,** 238 (1969).
2. McCarthy, D. W. J. & Probst, R. G. Detection of an infrared source near VB 8: the first extra-solar planet? *Bull. Am. Astron. Soc.* **16,** 965 (1984).
3. Latham, D. W., Stefanik, R. P., Mazeh, T., Mayor, M. & Burki, G. The unseen companion of HD114762 — a probable brown dwarf. *Nature* **339,** 38–40 (1989).
4. Wolszczan, A. & Frail, D. A. A planetary system around the millisecond pulsar PSR1257 + 12. *Nature* **355,** 145–147 (1992).
5. Mayor, M. & Queloz, D. A Jupiter-mass companion to a solar-type star. *Nature* **378,** 355–359 (1995).
   This article reports the discovery of the first giant planet around a Sun-like star.
6. Barge, P. *et al.* in *The CoRoT Mission, Pre-Launch Status, Stellar Seismology and Planet Finding* (eds Fridlund, M. *et al.*) (ESA, 2006).
7. Borucki, W. *et al.* KEPLER: search for Earth-size planets in the habitable zone. in *Proc. IAU Symposium No. IAU253* (eds Pont, F., Sasselov, D. & Holman, M.) 289–299 (2009).
8. Griffin, R. F. A photoelectric radial-velocity spectrometer. *Astrophys. J.* **148,** 465 (1967).

9. Baranne, A., Mayor, M. & Poncet, J. L. Coravel — a new tool for radial velocity measurements. *Vistas Astron.* **23,** 279–316 (1979).
10. Campbell, B. & Walker, G. A. H. Precision radial velocities with an absorption cell. *Publ. Astron. Soc. Pacif.* **91,** 540–545 (1979).
This paper describes the use of absorption cells to perform high-precision Doppler measurement.
11. Walker, G. A. H. in *Complementary Approaches to Double and Multiple Star Research* (eds McAlister, H. A. & Hartkopf, W. I.) 67 (ASP, 1992).
12. Vogt, S. S. *et al.* HIRES: the high-resolution echelle spectrometer on the Keck 10-m Telescope. *Proc. SPIE* **2198,** 362 (1994).
13. Baranne, A. *et al.* ELODIE: a spectrograph for accurate radial velocity measurements. *Astron. Astrophys. Suppl. Ser.* **119,** 373–390 (1996).
14. Butler, R. P. *et al.* Attaining Doppler precision of 3 m s^{-1}. *Publ. Astron. Soc. Pacif.* **108,** 500 (1996).
15. Mayor, M. & Udry, S. Mass function and distributions of the orbital elements of substellar companions. *ASP Conf. Ser.* **219,** 441 (2000).
16. Santos, N. C. *et al.* The HARPS survey for southern extra-solar planets II. A 14 Earth-masses exoplanet around μ Arae. *Astron. Astrophys.* **426,** L19–L23 (2004).
The discovery of the first sub-Neptune-mass planet is reported in this paper.
17. Mayor, M. *et al.* Setting new standards with HARPS. *Messenger* **114,** 20–24 (2003).
18. Dumusque, X. *et al.* An Earth-mass planet orbiting α Centauri B. *Nature* **491,** 207–211 (2012).
This article describes the detection of the lowest-ever measure radial-velocity signal induced by a planetary companion.
19. Pepe, F. A. & Lovis, C. From HARPS to CODEX: exploring the limits of Doppler measurements. *Phys. Scr.* **130,** 014007 (2008).
This paper gives an overview of the limiting factors of Doppler velocimetry and how they may be overcome.
20. Lovis, C., Mayor, M., Pepe, F., Queloz, D. & Udry, S. Pushing down the limits of RV precision with HARPS. *Astron. Soc. Pacif. Conf. Ser.* **398,** 455 (2008).
21. Bouchy, F., Pepe, F. & Queloz, D. Fundamental photon noise limit to radial velocity measurements. *Astron. Astrophys.* **374,** 733–739 (2001).
22. Perruchot, S. *et al.* Higher-precision radial velocity measurements with the SOPHIE spectrograph using octagonal-section fibers. *Proc. SPIE* **8151,** 815115 (2011).
23. Chazelas, B. Study of optical fiber scrambling to improve radial velocity measurements: simulations and experiments. Available at http://exoplanets.astro.psu.edu/workshop/program.html (2010).
24. Pepe, F., Mayor, M., Queloz, D. & Udry, S. Towards 1 ms-1 RV accuracy. *Proc. IAU Symp.* **202,** 103 (2004).
25. Saar, S. H. & Fischer, D. Correcting radial velocities for long-term magnetic activity variations. *Astrophys. J.* **534,** L105–L108 (2000).
26. Santos, N. C. *et al.* The CORALIE survey for Southern extra-solar planets. IV. Intrinsic stellar limitations to planet searches with radial-velocity techniques. *Astron. Astrophys.* **361,** 265–272 (2000).
27. Narayan, R., Cumming, A. & Lin, D. N. C. Radial velocity detectability of low-mass extrasolar planets in close orbits. *Astrophys. J.* **620,** 1002–1009 (2005).
28. Wright, J. T. Radial velocity jitter in stars from the California and Carnegie planet search at Keck observatory. *Publ. Astron. Soc. Pacif.* **117,** 657–664 (2005).
29. Reiners, A. Activity-induced radial velocity jitter in a flaring M dwarf. *Astron. Astrophys.* **498,** 853–861 (2009).
30. Lagrange, A. M., Meunier, N., Desort, M. & Malbet, F. Using the Sun to estimate Earth-like planets detection capabilities. III. Impact of spots and plages on astrometric detection. *Astron. Astrophys.* **528,** L9 (2011).
31. Martínez-Arnáiz, R., Maldonado, J., Montes, D., Eiroa, C. & Montesinos, B. Chromospheric activity and rotation of FGK stars in the solar vicinity. An estimation of the radial velocity jitter. *Astron. Astrophys.* **520,** A79 (2010).
32. Boisse, I. *et al.* Disentangling between stellar activity and planetary signals. *Astron. Astrophys.* **528,** A4 (2011).
33. Dumusque, X., Santos, N. C., Udry, S., Lovis, C. & Bonfils, X. Stellar noise and planet detection. in *Proc. 276th IAU Symposium* 527–529 (2011).
34. Dumusque, X., Lovis, C., Udry, S. & Santos, N. C. Stellar noise and planet detection. II. Radial-velocity noise induced by magnetic cycles. in *Proc. 276th IAU Symposium* 530–532 (2011).
35. Cegla, H. M. *et al.* Stellar jitter from variable gravitational redshift: implications for radial velocity confirmation of habitable exoplanets. *Mon. Not. R. Astron. Soc.* **421,** L54–L58 (2012).
36. Gettel, S. *et al. Correcting Astrophysical Noise in HARPS-N RV Measurements.* Available at http://www.mpia-hd.mpg.de/homes/ppvi/posters/2K024.html (2013).
37. Cegla, H. M., Stassun, K. G., Watson, C. A., Bastien, F. A. & Pepper, J. Estimating stellar radial velocity variability from Kepler and GALEX: implications for the radial velocity confirmation of exoplanets. *Astrophys. J.* **780,** 104 (2014).
38. Bastien, F. A. *et al.* Radial velocity variations of photometrically quiet, chromospherically inactive Kepler stars: a link between RV jitter and photometric flicker. *Astron. J.* **147,** 29 (2014).
39. Barnes, J. R. *et al.* Precision radial velocities of 15 M5–M9 dwarfs. *Mon. Not. R. Astron. Soc.* **439,** 3094–3113 (2014).
40. Pepe, F., Mayor, M. & Rupprecht, G. HARPS: ESO's coming planet searcher. Chasing exoplanets with the La Silla 3.6-m telescope. *Messenger* **110,** 9–14 (2002).
41. Pasquini, L., Cristiani, S., Garcia-Lopez, R., Haehnelt, M. & Mayor, M. CODEX: an ultra-stable high resolution spectrograph for the E-ELT. *Messenger* **140,** 20–21 (2010).
42. Pepe, F. *et al.* ESPRESSO: the next European exoplanet hunter. *Astron. Nachr.* **335,** 8 (2014).
43. Szentgyorgyi, A. *et al.* The GMT-CfA, Carnegie, Catolica, Chicago Large Earth Finder (G-CLEF): a general purpose optical echelle spectrograph for the GMT with precision radial velocity capability. *Proc. SPIE* **8446,** 84461H (2012).
44. Kasting, J. F., Whitmire, D. P. & Reynolds, R. T. Habitable zones around main sequence stars. *Icarus* **101,** 108 (1993).
This paper provides a definition for the habitable zone.
45. Heacox, W. D. in *Fiber Optics in Astronomy* (ed. Barden, S. C.) 204–235 (*Astron. Soc. Pacif.*, 1988)
46. Hubbard, E. N., Angel, J. R. P. & Gresham, M. S. Operation of a long fused silica fiber as a link between telescope and spectrograph. *Astrophys. J.* **229,** 1074–1078 (1979).
47. Barden, S. C., Ramsey, L. W. & Truax, R. J. Evaluation of some fiber optical waveguides for astronomical instrumentation. *Publ. Astron. Soc. Pacif.* **93,** 154–162 (1981).
48. Heacox, W. D. & Connes, P. Optical fibers in astronomical instruments. *Astron. Astrophys.* **3,** 169–199 (1992).
This paper contains a complete and detailed discussion of instrumental effects and in particular the use of optical fibres for image scrabbling.
49. Hunter, T. R. & Ramsey, L. W. Scrambling properties of optical fibers and the performance of a double scrambler. *Publ. Astron. Soc. Pacif.* **104,** 1244–1251 (1992).
50. Avila, G., Buzzoni, B. & Casse, M. Fiber characterization and compact scramblers at ESO. *Proc. SPIE* **3355,** 900–904 (1998).
51. Avila, G. & Singh, P. Optical fiber scrambling and light pipes for high accuracy radial velocities measurements. *Proc. SPIE* **7018,** 70184W (2008).
52. Chazelas, B., Pepe, F. & Wildi, F. Optical fibers for precise radial velocities: an update. *Proc. SPIE* **8450,** 845013 (2012).
53. Plavchan, P. P. *et al.* Precision near-infrared radial velocity instrumentation II: noncircular core fiber scrambler. *Proc. SPIE* **8864,** 88640G (2013).
54. Cosentino, R. *et al.* Harps-N: the new planet hunter at TNG. *Proc. SPIE* **8446,** 84461V (2012).
55. Bouchy, F. *et al.* SOPHIE+: first results of an octagonal-section fiber for high-precision radial velocity measurements. *Astron. Astrophys.* **549,** 49 (2013).
56. Mahadevan, S. & Ge, J. The use of absorption cells as a wavelength reference for precision radial velocity measurements in the near-infrared. *Astrophys. J.* **692,** 1590–1596 (2009).
57. Osterman, S. *et al.* A proposed laser frequency comb-based wavelength reference for high-resolution spectroscopy. *Proc. SPIE* **6693,** 66931G (2007).
58. Li, C.-H. *et al.* A laser frequency comb that enables radial velocity measurements with a precision of 1cm s^{-1}. *Nature* **452,** 610–612 (2008).
59. Steinmetz, T. *et al.* Laser frequency combs for astronomical observations. *Science* **321,** 1335–1337 (2008).
60. Schettino, G. *et al.* The Astro-Comb project. *Proc. SPIE* **7808,** 78081Q (2010).
61. Phillips, D. F. *et al.* Calibration of an echelle spectrograph with an astro-comb: a laser frequency comb with very high repetition rate. *Proc. SPIE* **8446,** 844680 (2012).
62. Johnson, A. R. *et al.* Microresonator-based comb generation without an external laser source. *Opt. Express* **22,** 1394 (2014).
63. Wildi, F., Pepe, F. & Chazelas, B. Curto, Lo, G. & Lovis, C. The performance of the new Fabry-Perot calibration system of the radial velocity spectrograph HARPS. *Proc. SPIE* **8151,** 81511F (2011).
64. Schäfer, S. & Reiners, A. Two Fabry-Perot interferometers for high precision wavelength calibration in the near-infrared. *Proc. SPIE* **8446,** 844694 (2012).
65. Halverson, S. *et al.* Development of fiber Fabry-Perot interferometers as stable near-infrared calibration sources for high resolution spectrographs. *Publ. Astron. Soc. Pacif.* **126,** 445–458 (2014).
66. Schwab, C. *et al.* Stabilizing a Fabry-Perot etalon to 3 cm/s for spectrograph calibration. Preprint at http://arxiv.org/abs/1404.0004 (2014).
67. Reiners, A. *et al.* Detecting planets around very low mass stars with the radial velocity method. *Astrophys. J.* **710,** 432–443 (2010).
68. Desort, M., Lagrange, A. M., Galland, F., Udry, S. & Mayor, M. Search for exoplanets with the radial-velocity technique: quantitative diagnostics of stellar activity. *Astron. Astrophys.* **473,** 983–993 (2007).
69. Huélamo, N. *et al.* TW Hydrae: evidence of stellar spots instead of a hot Jupiter. *Astron. Astrophys.* **489,** L9–L13 (2008).
70. Barnes, J. R., Jeffers, S. V. & Jones, H. R. A. The effect of M dwarf starspot activity on low-mass planet detection thresholds. *Mon. Not. R. Astron. Soc.* **412,** 1599–1610 (2011).
71. Oliva, E. *et al.* The GIANO spectrometer: towards its first light at the TNG. *Proc. SPIE* **8446,** 84463T (2012).
72. Quirrenbach, A. *et al.* CARMENES. I: instrument and survey overview. *Proc. SPIE* **8446,** 84460R (2012).
73. Tamura, M. *et al.* Infrared Doppler instrument for the Subaru Telescope (IRD). *Proc. SPIE* **8446,** 84461T (2012).
74. Mahadevan, S. *et al.* The habitable-zone planet finder: a stabilized fiber-fed NIR spectrograph for the Hobby-Eberly Telescope. *Proc. SPIE* **8446,** 84461S (2012).
75. Ge, J. *et al.* High resolution Florida IR silicon immersion grating spectrometer and an M dwarf planet survey. *Proc. SPIE* **8446,** 844630 (2012).
76. Delfosse, X. *et al.* World-leading science with SPIRou – the nIR spectropolarimeter/high-precision velocimeter for CFHT. *Proc. SF2A* 497–508 (2013).
77. Schwab, C., Leon-Saval, S. G., Betters, C. H., Bland-Hawthorn, J. & Mahadevan, S. Single mode, extreme precision Doppler spectrographs. *IAU Proc.* Preprint at http://arxiv.org/abs/1212.4867 (2012).
78. Wright, J. T. *et al.* The frequency of hot Jupiters orbiting nearby solar-type stars. *Astrophys. J.* **753,** 160 (2012).

79. Collier Cameron, A. *et al*. WASP-1b and WASP-2b: two new transiting exoplanets detected with SuperWASP and SOPHIE. *Mon. Not. R. Astron. Soc.* **375,** 951–957 (2007)
80. Bakos, G. Á. et al. HAT-P-1b: a large-radius, low-density exoplanet transiting one member of a stellar binary. *Astrophys. J.* **656,** 552–559 (2007).
81. Alonso, R. *et al*. TrES-1: the transiting planet of a bright K0 V star. *Astrophys. J.* **613,** L153 (2004).
82. Udalski, A. *et al*. The optical gravitational lensing experiment. Search for planetary and low-luminosity object transits in the galactic disk. Results of 2001 Campaign — Supplement. *Acta Astron*. **52,** 115–128 (2002).
83. Konacki, M., Torres, G., Jha, S. & Sasselov, D. D. An extrasolar planet that transits the disk of its parent star. *Nature* **421,** 507–509 (2003).
This paper reports the first discovery of an exoplanet with a ground-based transit survey.
84. Moutou, C. *et al*. CoRoT: harvest of the exoplanet program. *Icarus* **226,** 1625–1634 (2013).
85. Charbonneau, D., Irwin, J., Nutzman, P. & Falco, E. E. The MEarth project to detect habitable super Earth exoplanets. *Bull. Am. Astron. Soc*. **40,** 242 (2008).
86. Charbonneau, D. *et al*. A super-Earth transiting a nearby low-mass star. *Nature* **462,** 891–894 (2009).
87. Bean, J. L., Miller-Ricci Kempton, E. M.-R. & Homeier, D. A ground-based transmission spectrum of the super-Earth exoplanet GJ 1214b. *Nature* **468,** 669–672 (2010).
88. Croll, B. *et al*. Broadband transmission spectroscopy of the super-Earth GJ 1214b suggests a low mean molecular weight atmosphere. *Astrophys. J.* **736,** 78 (2011).
89. Crossfield, I. J. M., Barman, T. & Hansen, B. M. S. High-resolution, differential, near-infrared transmission spectroscopy of GJ 1214b. *Astrophys. J.* **736,** 132 (2011).
90. Kreidberg, L., Bean, J. L., Désert, J.-M. & Benneke, B. Clouds in the atmosphere of the super-Earth exoplanet GJ 1214b. *Nature* **505,** 69–72 (2014).
This paper presents a high-precision medium-resolution transmission spectrum of a super-Earth.
91. Désert, J.-M. *et al*. Observational evidence for a metal rich atmosphere on the super-Earth GJ1214b. *Astrophys. Lett.* **731,** L40 (2011).
92. Charbonneau, D., Brown, T. M., Latham, D. W. & Mayor, M. Detection of planetary transits across a Sun-like star. *Astrophys. J.* **529,** L45 (2000).
93. Henry, G. W., Marcy, G. W., Butler, R. P. & Vogt, S. S. A. Transiting 51 Peg-like planet. *Astrophys. J.* **529,** L41 (2000).
94. Brown, T. M., Charbonneau, D., Gilliland, R. L., Noyes, R. W. & Burrows, A. Hubble space telescope time-series photometry of the transiting planet of HD 209458. *Astrophys. J.* **552,** 699 (2001).
This paper reports on the first observation of an exoplanet transit from space, with the Hubble Space Telescope.
95. Woodgate, B. E. *et al*. The space telescope imaging spectrograph design. *Publ. Astron. Soc. Pacif.* **110,** 1183 (1998).
96. Charbonneau, D., Brown, T. M., Noyes, R. W. & Gilliland, R. L. Detection of an extrasolar planet atmosphere. *Astrophys. J.* **568,** 377 (2002).
97. Vidal-Madjar, A. *et al*. An extended upper atmosphere around the extrasolar planet HD 209458b. *Nature* **422,** 143–146 (2003).
98. Pont, F. *et al*. Hubble Space Telescope time-series photometry of the planetary transit of HD 189733: no moon, no rings, starspots. *Astron. Astrophys*. **476,** 1347 (2007).
99. Fazio, G. G. *et al*. The Infrared Array Camera (IRAC) for the Spitzer Space Telescope. *Astrophys. J. Suppl. Ser*. **154,** 10 (2004).
100. Charbonneau, D. *et al*. Detection of thermal emission from an extrasolar planet. *Astrophys. J.* **626,** 523 (2005).
101. Deming, D., Seager, S., Richardson, L. J. & Harrington, J. Infrared radiation from an extrasolar planet. *Nature* **434,** 740–743 (2005).
102. Knutson, H. A. *et al*. A map of the day–night contrast of the extrasolar planet HD 189733b. *Nature* **447,** 183–186 (2007).
103. Richardson, L. J., Harrington, J., Seager, S. & Deming, D. A Spitzer infrared radius for the transiting extrasolar planet HD 209458b. *Astrophys. J.* **649,** 1043 (2006).
104. Ehrenreich, D. *et al*. A Spitzer search for water in the transiting exoplanet HD 189733b. *Astrophys. J.* **668,** L179 (2007).
105. Beaulieu, J. P., Carey, S., Ribas, I. & Tinetti, G. Primary transit of the planet HD 189733b at 3.6 and 5.8 μm. *Astrophys. J.* **677,** 1343 (2008).
106. Désert, J.-M. *et al*. Search for carbon monoxide in the atmosphere of the transiting exoplanet HD 189733b. *Astrophys. J.* **699,** 478 (2009).
107. Thompson, R. I. NICMOS: the next U.S. infrared space mission. *Proc. SPIE* **2198,** 1202–1213 (1994).
108. Swain, M. R., Vasisht, G. & Tinetti, G. The presence of methane in the atmosphere of an extrasolar planet. *Nature* **452,** 329 (2008).
109. Sing, D. K. *et al*. Transit spectrophotometry of the exoplanet HD 189733b. I. Searching for water but finding haze with HST NICMOS. *Astron. Astrophys*. **505,** 891–899 (2009).
110. Gibson, N. P., Pont, F. & Aigrain, S. A new look at NICMOS transmission spectroscopy of HD 189733, GJ-436 and XO-1: no conclusive evidence for molecular features. *Mon. Not. R. Astron. Soc.* **411,** 2199 (2011).
This paper explores the impact of instrumental systematics in transmission spectroscopy, which can lead to false-positive detections.
111. Houck, J. R. *et al*. The Infrared Spectrograph (IRS) on the Spitzer Space Telescope. *Astrophys. J. Suppl. Ser*. **154,** 18–24 (2004).
112. Grillmair, C. J. *et al*. A Spitzer spectrum of the exoplanet HD 189733b. *Astrophys. J.* **658,** L115 (2007).
113. Grillmair, C. J. *et al*. Strong water absorption in the dayside emission spectrum of the planet HD 189733b. *Nature* **456,** 767–769 (2008).
114. Demory, B.-O. *et al*. Detection of a transit of the super-Earth 55 Cancri e with warm Spitzer. *Astron. Astrophys*. **533,** 114 (2011).
115. Van Grootel, V. *et al*. Transit confirmation and improved stellar and planet parameters for the super-Earth HD 97658 b and its host star. *Astrophys. J.* **796,** 2 (2014).
116. Snellen, I. A. G., Albrecht, S., de Mooij, E. J. W. & Le Poole, R. S. Ground-based detection of sodium in the transmission spectrum of exoplanet HD 209458b. *Astron. Astrophys*. **487,** 357 (2008).
This paper reports the first detection of an exoplanetary atmosphere from the ground, with a high-resolution spectrograph installed at the Subaru telescope.
117. Narita, N. *et al*. Subaru HDS transmission spectroscopy of the transiting extrasolar planet HD 209458b. *Publ. Astron. Soc. Jpn* **57,** 471 (2005).
118. Sing, D. K. *et al*. Gran Telescopio Canarias OSIRIS transiting exoplanet atmospheric survey: detection of potassium in XO-2b from narrowband spectrophotometry. *Astron. Astrophys*. **527,** 73 (2011).
119. Colon, K. D. *et al*. Measuring potassium in exoplanet atmospheres with the OSIRIS tunable filter. *RevMexAA (Serie de Conferencias)* **42,** 1–2 (2013).
120. Snellen, I. A. G., de Kok, R. J., de Mooij, E. J. W. & Albrecht, S. The orbital motion, absolute mass and high-altitude winds of exoplanet HD209458b. *Nature* **465,** 1049–1051 (2010).
121. Birkby, J. L. *et al*. Detection of water absorption in the dayside atmosphere of HD 189733 b using ground-based high-resolution spectroscopy at 3.2 microns. *Mon. Not. R. Astron. Soc*. **436,** L35 (2013).
122. Lockwood, A. C., Johnson, J. A. & Bender, C. F. Near-IR direct detection of water vapor in τ Boötis b. *Astrophys. Lett*. **783,** L29 (2014).
123. Brogi, M. *et al*. The signature of orbital motion from the dayside of the planet τ Boötis b. *Nature* **486,** 502–504 (2012).
124. Lagrange, A.-M. *et al*. A probable giant planet imaged in the β Pictoris disk. VLT/NaCo deep L′-band imaging. *Astron. Astrophys*. **493,** L21 (2009).
This paper reports the near-infrared detection of the giant exoplanet β Pictoris b from adaptive-optics high-contrast observations with the NACO instrument on the VLT.
125. Snellen, I. *et al*. The fast spin-rotation of the young extrasolar planet β Pictoris b. *Nature* **509,** 63–65 (2014).
This paper details how ground-based, high-resolution spectroscopy can be used to probe the atmosphere of a directly imaged exoplanet and measure its spin-rotation.
126. Green, J. C. Cosmic origins spectrograph. *Proc. SPIE* **4498,** 229–238 (2001).
127. Kimble, R. A., MacKenty, J. W., O'Connell, R. W. & Townsend, J. A. Wide Field Camera 3: a powerful new imager for the Hubble Space Telescope. *Proc. SPIE* **7010,** 70101E (2008).
128. Linsky, J. L. *et al*. Observations of mass loss from the transiting exoplanet HD 209458b. *Astrophys. J.* **717,** 1291 (2010).
129. Lecavelier des Etangs, A. *et al*. Temporal variations in the evaporating atmosphere of the exoplanet HD 189733b. *Astron. Astrophys*. **543,** L4 (2012).
130. Bourrier, V. *et al*. Atmospheric escape from HD 189733b observed in HI Lyman-α: detailed analysis of HST/STIS September 2011 observations. *Astron. Astrophys*. **551,** A63 (2013).
131. Ehrenreich, D. *et al*. Hint of a transiting extended atmosphere on 55 Cancri b. *Astron. Astrophys*. **547,** 18 (2012).
132. Fossati, L. *et al*. Metals in the exosphere of the highly irradiated planet WASP-12b. *Astrophys. Lett.* **714,** L222 (2010).
133. Kulow, J. R., France, K., Linsky, J. & Loyd, R. O. P. Lyα Transit spectroscopy and the neutral hydrogen tail of the hot Neptune GJ 436b. *Astrophys. J.* **786,** 132 (2014).
134. Huitson, C. M., Sing, D. K., Vidal-Madjar, A., Ballester, G. E. & Lecavelier des Etangs, A. Temperature-pressure profile of the hot Jupiter HD 189733b from HST sodium observations: detection of upper atmospheric heating. *Mon. Not. R. Astron. Soc.* **422,** 2477 (2012).
135. Evans, T. M. *et al*. The deep blue color of HD 189733b: albedo measurements with Hubble Space Telescope/Space Telescope imaging spectrograph at visible wavelengths. *Astrophys. Lett.* **772,** L16 (2013).
136. Berta, Z. K. *et al*. The flat transmission spectrum of the super-Earth GJ1214b from Wide Field Camera 3 on the Hubble Space Telescope. *Astrophys. J.* **747,** 35 (2012).
137. Deming, D. *et al*. Infrared transmission spectroscopy of the exoplanets HD 209458b and XO-1b using the Wide Field Camera-3 on the Hubble Space Telescope. *Astrophys. J.* **774,** 95 (2013).
138. Mandell, A. *et al*. Exoplanet transit spectroscopy using WFC3: WASP-12 b, WASP-17 b, and WASP-19 b. *Astrophys. J.* **779,** 128 (2013).
139. Knutson, H. A., Benneke, B., Deming, D. & Homeier, D. A featureless transmission spectrum for the Neptune-mass exoplanet GJ 436b. *Nature* **505,** 66–68 (2014).
140. Ranjan, S. *et al*. Atmospheric characterization of 5 hot Jupiters with Wide Field Camera 3 on the Hubble Space Telescope. *Astrophys. J.* **785,** 148 (2014).
141. Gibson, N. P. *et al*. Probing the haze in the atmosphere of HD 189733b with HST/WFC3 transmission spectroscopy. *Mon. Not. R. Astron. Soc*. **422,** 753 (2012).
142. Knutson, H. A. *et al*. Hubble Space Telescope near-IR transmission spectroscopy of the super-Earth HD 97658b. Preprint at http://arxiv.org/abs/1403.4602 (2014).
143. Ehrenreich, D. *et al*. Near-infrared transmission spectrum of the warm-Uranus GJ 3470b with the Wide Field Camera-3 on the Hubble Space Telescope. Preprint at http://arxiv.org/abs/1405.1056 (2014).

144. Marois, C. *et al*. Direct imaging of multiple planets orbiting the star HR 8799. *Science* **322,** 1348–1352 (2008).
This paper reports on the discovery of three planetary companions, directly imaged in the near-infrared at two different epochs with the adaptive-optics systems at the Gemini and Keck telescopes, and the corresponding facility near-infrared cameras.
145. Kalas, P. *et al*. Optical images of an exosolar planet 25 light-years from Earth. *Science* **322,** 1345–1348 (2008).
146. Kalas, P., Graham, J. R., Fitzgerald, M. P. & Clampin, M. HST/STIS imaging of Fomalhaut: new main belt structure and confirmation of Fomalhaut b's eccentric orbit. *Proc. IAU* **8,** 204–207 (2014).
147. Lagrange, A.-M. *et al*. Constraining the orbit of the possible companion to β Pictoris. New deep imaging observations. *Astron. Astrophys.* **506,** 927 (2009).
148. Lagrange, A.-M. *et al*. A giant planet imaged in the disk of the young star β Pictoris. *Science* **329,** 57 (2010).
149. Quanz, S. P. *et al*. First results from Very Large Telescope NACO apodizing phase plate: 4 μm images of the exoplanet β Pictoris b. *Astrophys. Lett.* **722,** L49 (2010).
150. Smith, B. A. & Terrile, R. J. A circumstellar disk around β Pictoris. *Science* **226,** 1421–1424 (1984).
151. Chauvin, G. *et al*. A giant planet candidate near a young brown dwarf. *Astron. Astrophys.* **425,** L29–L32 (2004).
This paper reports the first direct detection of a planetary-mass object, orbiting a brown dwarf, with the NACO instrument on the VLT.
152. Chabrier, G., Johansen, A., Janson, M. & Rafikov, R. Giant planet and brown dwarf formation. Preprint at http://arxiv.org/abs/1401.7559 (2014).
153. Davies, R. & Kasper, M. Adaptive optics for astronomy. *Annu. Rev. Astron. Astrophys.* **50,** 305–351 (2012).
154. Nielsen, E. L. & Close, L. M. A uniform analysis of 118 stars with high-contrast imaging: long-period extrasolar giant planets are rare around Sun-like stars. *Astrophys. J.* **717,** 878–896 (2010).
155. Fusco, T. *et al*. Integration of SAXO, the VLT-SPHERE extreme AO: final performance. *Proc. Third AO4ELT Conf*. http://dx.doi.org/10.12839/AO4ELT3.13327 (2013).
156. Macintosh, B. *et al*. The Gemini Planet Imager: first light. *Proc. Natl Acad. Sci. USA.* Preprint at http://arxiv.org/abs/1403.7520 (2014).
157. Close, L. *et al*. Into the blue: AO science in the visible with MagAO. *Proc. Third AO4ELT Conf*. http://dx.doi.org/10.12839/AO4ELT3.13387 (2013).
158. Marois, C., Lafrenière, D., Doyon, R., Macintosh, B. & Nadeau, D. Angular differential imaging: a powerful high-contrast imaging technique. *Astrophys. J.* **641,** 556 (2006).
159. Amara, A. & Quanz, S. P. PYNPOINT: an image processing package for finding exoplanets. *Mon. Not. R. Astron. Soc.* **427,** 948–955 (2012).
160. Quanz, S. P. *et al*. Resolving the inner regions of circumstellar discs with VLT/NACO polarimetric differential imaging. *Messenger* **146,** 25–27 (2011).
161. Milli, J. *et al*. Prospects of detecting the polarimetric signature of the Earth-mass planet α Centauri B b with SPHERE/ZIMPOL. *Astron. Astrophys.* **556,** 64 (2013).
162. Crepp, J. R. *et al*. Speckle suppression with the Project 1640 Integral Field Spectrograph. *Astrophys. J.* **729,** 132 (2011).
163. Ireland, M. J. Phase errors in diffraction-limited imaging: contrast limits for sparse aperture masking. *Mon. Not. R. Astron. Soc.* **433,** 1718–1728 (2013).
164. Kenworthy, M. A. *et al*. First on-sky high-contrast imaging with an apodizing phase plate. *Astrophys. J.* **660,** 762–769 (2007).
165. Mawet, D. *et al*. L'-band AGPM vector vortex coronagraph's first light on VLT/NACO. Discovery of a late-type companion at two beamwidths from an F0V star. *Astron. Astrophys.* **552,** L13 (2013).
166. Guyon, O. Phase-induced amplitude apodization of telescope pupils for extrasolar terrestrial planet imaging. *Astron. Astrophys.* **404,** 379–387 (2003).
167. Mawet, D. *et al*. Review of small-angle coronagraphic techniques in the wake of ground-based second-generation adaptive optics systems. *Proc. SPIE* **8442,** 844204 (2012).
168. Bailey, V. *et al*. The large binocular telescope interferometer and adaptive optics system: on-sky performance and results. *Proc. IAU* **8,** 26–27 (2014).
169. Codona, J. L. & Kenworthy, M. Focal plane wavefront sensing using residual adaptive optics speckles. *Astrophys. J.* **767,** 100 (2013).
170. Quanz, S. P., Crossfield, I., Meyer, M. R., Schmalzl, E. & Held, J. Direct detection of exoplanets in the 3–10 micron range with E-ELT/METIS. *Int. J. Astrobiol*. Preprint at http://arxiv.org/abs/1404.0831 (2014).
171. Sozzetti, A. *et al*. Astrometric detection of giant planets around nearby M dwarfs: the Gaia potential. *Mon. Not. R. Astron. Soc.* **437,** 497–509 (2014).
172. Sahlmann, J. *et al*. Narrow-angle astrometry with PRIMA. *Proc. SPIE* **8445,** 84450S1 (2012).
173. Howell, S. B. *et al*. The K2 mission: characterization and early results. *Publ. Astron. Soc. Pacif*. Preprint at http://arxiv.org/abs/1402.5163 (2014).
174. Ricker, G. R. *et al*. The Transiting Exoplanet Survey Satellite (TESS). *Bull. Am. Astron. Soc.* **41,** 193 (2009).
175. Rauer, H. *et al*. The PLATO 2.0 mission. *Exp. Astron*. Preprint at http://arxiv.org/abs/1310.0696 (2013).
176. Snellen, I. *et al*. Ground-based search for the brightest transiting planets with Multi-site All Sky Camera - MASCARA. *Proc. SPIE* **8444,** 844401 (2012)
177. CHEOPS Study Team. *CHEOPS Definition Study Report (Red Book)* (ESA, 2013).
178. Broeg, C. *et al*. CHEOPS: A transit photometry mission for ESA's small mission programme. *EPJ Web Conf*. **47,** 03005 (2013).
179. Snellen, I. A. G., de Kok, R. J., le Poole, R., Brogi, M. & Birkby, J. Finding extraterrestrial life using ground-based high-dispersion spectroscopy. *Astrophys. J.* **764,** 182 (2013).
180. Vogt, S. S. The Lick Observatory Hamilton Echelle Spectrometer. *Publ. Astron. Soc. Pacif.* **99,** 1214–1228 (1987).
181. Horton, A. *et al*. CYCLOPS2: the fibre image slicer upgrade for the UCLES high resolution spectrograph. *Proc. SPIE* **8446,** 84463A (2012).
182. Dekker, H., D'Odorico, S., Kaufer, A., Delabre, B. & Kotzlowski, H. Design, construction, and performance of UVES, the echelle spectrograph for the UT2 Kueyen Telescope at the ESO Paranal Observatory. *Proc. SPIE* **4008,** 534 (2000).
183. Tull, R. G. High-resolution fiber-coupled spectrograph of the Hobby-Eberly Telescope. *Proc. SPIE* **3355,** 387 (1998).
184. Noguchi, K. *et al*. High Dispersion Spectrograph (HDS) for the Subaru Telescope. *Publ. Astron. Soc. Jpn*. **54,** 855–864 (2002).
185. Kaufer, A. *et al*. Commissioning FEROS, the new high-resolution spectrograph at La-Silla. *Messenger* **95,** 8–12 (1999).
186. Bernstein, R., Shectman, S. A., Gunnels, S. M., Mochnacki, S. & Athey, A. E. MIKE: a Double Echelle Spectrograph for the Magellan Telescopes at Las Campanas Observatory. *Proc. SPIE* **4841,** 1694 (2003).
187. Bouchy, F. & Team, S. in *Proc. Colloquium of the Tenth Anniversary of 51 Peg-b* 319–325 http://www.obs-hp.fr/www/pubs/Coll51Peg/proceedings.html (2006).
188. Kaeufl, H.-U. *et al*. CRIRES: a high-resolution infrared spectrograph for ESO's VLT. *Proc. SPIE* **5492,** 1218 (2004).
189. Crane, J. D. *et al*. The Carnegie Planet Finder Spectrograph: integration and commissioning. *Proc. SPIE* **7735,** 773553 (2010).
190. Chakraborty, A. *et al*. First light results from PARAS: the PRL Echelle Spectrograph. *Proc. SPIE* **7735,** 77354N (2010).
191. Aceituno, J. *et al*. CAFE: Calar Alto Fiber-fed Échelle spectrograph. *Astron. Astrophys.* **552,** 31 (2013).
192. Schwab, C., Spronck, J. F. P., Tokovinin, A. & Fischer, D. A. Design of the CHIRON high-resolution spectrometer at CTIO. *Proc. SPIE* **7735,** 77354G (2010).
193. Vogt, S. S. *et al*. APF - The Lick Observatory Automated Planet Finder. *Publ. Astron. Soc. Pacif.* Preprint at http://arxiv.org/abs/1402.6684 (2014).
194. Ge, J. *et al*. Design and performance of a new generation, compact, low cost, very high Doppler precision and resolution optical spectrograph. *Proc. SPIE* **8446,** 84468R (2012).
195. Bramall, D. G. *et al*. The SALT HRS spectrograph: instrument integration and laboratory test results. *Proc. SPIE* **8446,** 84460A (2012).
196. Strassmeier, K. G. *et al*. PEPSI: the Potsdam Echelle Polarimetric and Spectroscopic Instrument for the LBT. *Proc. SPIE* **7014,** 70140N (2008).
197. Brown, T. M., Noyes, R. W., Nisenson, P., Korzennik, S. G. & Horner, S. The AFOE: a spectrograph for precise Doppler studies. *Publ. Astron. Soc. Pacif.* **106,** 1285–1297 (1994).
198. Tull, R. G., MacQueen, P. J., Sneden, C. & Lambert, D. L. The high-resolution cross-dispersed echelle white-pupil spectrometer of the McDonald Observatory 2.7-m telescope. *Publ. Astron. Soc. Pacif.* **107,** 251–264 (1995).
199. Schneider, J., Dedieu, C., Le Sidaner, P., Savalle, R. & Zolotukhin, I. Defining and cataloging exoplanets: the exoplanet.eu database. *Astron. Astrophys.* **532,** A79 (2011).

Acknowledgements D.E. would like to dedicate this article to the memory of STIS Principal Investigator Bruce Woodgate who passed away in April 2014. This work has been carried out within the frame of the National Centre for Competence in Research 'PlanetS' supported by the Swiss National Science Foundation (SNSF). The authors acknowledge the financial support of the SNSF.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this paper at go.nature.com/azn8hn. Correspondence should be addressed to F.P. (Francesco.Pepe@unige.ch).

ARTICLE

doi:10.1038/nature13735

Multifunctional organoboron compounds for scalable natural product synthesis

Fanke Meng[1], Kevin P. McGrath[1] & Amir H. Hoveyda[1]

Efficient catalytic reactions that can generate C–C bonds enantioselectively, and ones that can produce trisubstituted alkenes diastereoselectively, are central to research in organic chemistry. Transformations that accomplish these two tasks simultaneously are in high demand, particularly if the catalysts, substrates and reagents are inexpensive and if the reaction conditions are mild. Here we report a facile multicomponent catalytic process that begins with a chemoselective, site-selective and diastereoselective copper–boron addition to a monosubstituted allene; the resulting boron-substituted organocopper intermediates then participate in a similarly selective allylic substitution. The products, which contain a stereogenic carbon centre, a monosubstituted alkene and an easily functionalizable *Z*-trisubstituted alkenylboron group, are obtained in up to 89 per cent yield, with more than 98 per cent branch-selectivity and stereoselectivity and an enantiomeric ratio greater than 99:1. The copper-based catalyst is derived from a robust heterocyclic salt that can be prepared in multigram quantities from inexpensive starting materials and without costly purification procedures. The utility of the approach is demonstrated through enantioselective synthesis of gram quantities of two natural products, namely rottnestol and herboxidiene (also known as GEX1A).

Enantioselective processes where a catalyst unites a pair of starting materials and then induces the resulting species to react with a third substrate are sought-after in chemistry[1,2]. Pathways that involve difficult-to-access intermediates and products then become feasible, and wasteful and costly procedures for isolation and/or purification of sensitive reagents become unnecessary[3]. Rare instances of such multicomponent processes can be found in phosphine–Ir or Ru-catalysed enantioselective reductive fusion of hydrogen, unsaturated hydrocarbons and carbonyl or imine compounds[4,5]. An unprecedented degree of complexity would result if a multitasking catalyst were to promote several transformations that are each selective on multiple levels, with the final product bearing the marks of every single discriminatory event; a representative pathway is shown in Fig. 1a.

Multicomponent synthesis of complex fragments

Boron-substituted alkenes are widely used multipurpose moieties. Single-catalyst/multisubstrate transformations that deliver multifunctional unsaturated organoboron compounds are therefore of great interest. In the first phase of our studies (Fig. 1b), we found that chemoselective addition of (phosphine)Cu–B(pin) (here B(pin) = (pinacolato)boron), derived from reaction of an *in situ* generated (phosphine)Cu–alkoxide with $B_2(pin)_2$, to a monosubstituted allene (versus an aldehyde) affords 2-B(pin)-substituted allylcopper complex **i**, which then reacts with an aldehyde (versus an allene) to afford homoallylic alkoxide **iii**. An assortment of aldol-type products were obtained after oxidative treatment in up to >99:1 diastereomeric ratio (d.r.) and 97:3 enantiomeric ratio (e.r.)[6]. In contrast, transformations with N-heterocyclic carbene (NHC) complexes of copper, while efficient, generated racemic products.

The above reactions give 1,1-disubstituted alkenylboron units because of a second-stage γ addition (**ii**), which causes the loss of an important attribute of the initially formed intermediate (**i**): a stereochemically defined and modifiable trisubstituted olefin. A multicomponent catalytic enantioselective process that preserves the trisubstituted alkenylboron group would have higher value. We thus envisioned a transformation involving chemo-, site- and stereoselective Cu–B(pin) addition to an allenyl substrate followed by chemo- and site-selective (branched versus linear) cross-coupling of the resulting allylcopper species through enantioselective allylic substitution (EAS). The envisioned catalytic sequence would furnish multifunctional organoboron products **v** by a single operation; this would be in contrast to the existing strategies where each functional unit must be installed individually through extended and less efficient sequences[7,8] (for a complete bibliography, see Supplementary Information). Such a process would be a significant addition to an important but limited group of catalytic allyl–allyl reactions. Site- and enantioselective incorporation of allyl groups through catalytic EAS has been confined to simple fragments introduced via allylboron[9], allylmagnesium[10] or allylic alcohol[11] compounds (see Supplementary Information for a complete bibliography).

The expected organoboron products (**v**, Fig. 1b) are rich in adaptable moieties. A stereogenic centre could be formed in the homoallylic position of a stereochemically defined trisubstituted alkenylboron unit that may be converted to other *E*- or *Z*-trisubstituted olefins. For instance, conversion of the C–B(pin) of **v** to a C–C bond with inversion of stereochemistry would deliver **vi**, which is a functional group found in numerous biologically active molecules; a notable case corresponds to a segment of immunosuppressive agent FK-506[12] (compare highlighted fragment in Fig. 1c). Efficient and stereoselective synthesis of such trisubstituted olefin-containing fragments remains a difficult problem. In previous efforts either the undesired *Z* olefin was removed from a near-equal mixture of isomers[13,14], or modification of a terminal alkyne by relatively lengthy routes was required[15]. The terminal olefin of the products is an asset as well: it would provide the opportunity for many types of modifications. One example entails conversion to an *E*,*E*-diene by sequential catalytic cross-metathesis with vinyl–B(pin)[16] and cross-coupling (Fig. 1c)[17], generating a fragment that is common to several biologically active natural products. The highlighted segments in nafuredin (NADH-fumarate reductase inhibitor[18,19]), milbemycin β_3 (insecticidal[20]), rottnestol (member of a family of antibiotics[21]) and herboxidiene (phytotoxic, anti-tumour[22]) are representative.

[1]Department of Chemistry, Merkert Chemistry Center, Boston College, Chestnut Hill, Massachusetts 02467, USA.

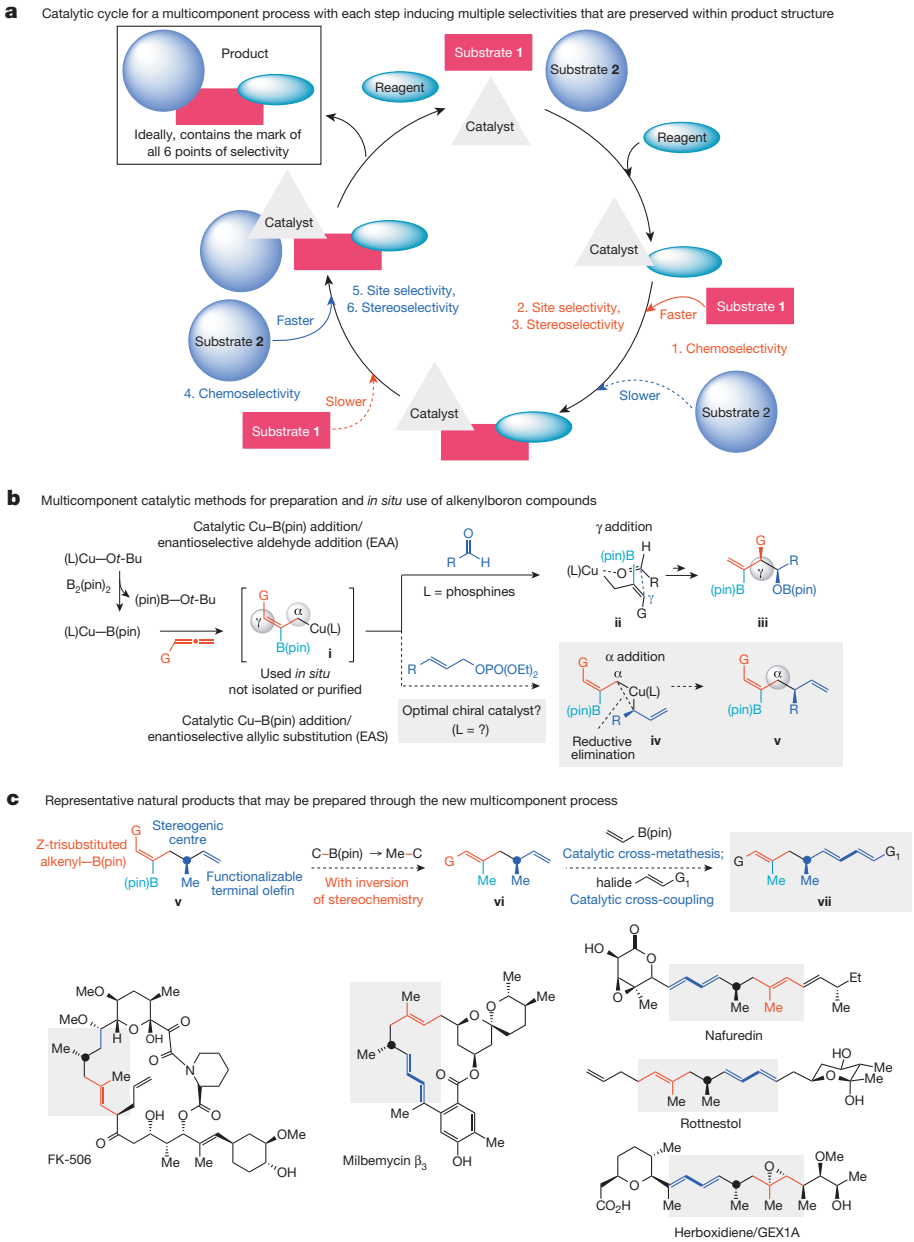


Figure 1 | Multicomponent catalytic enantioselective generation of alkenylboron compounds. **a**, The general scheme for a multicomponent catalytic cycle involving a reagent and two substrates might be envisioned to proceed by a sequence entailing multiple selectivity issues. Ideally, all points of selectivity would be retained. **b**, Catalytic stereoselective generation of an alkenyl–B(pin) intermediate (**i**), which might react *in situ* site-, diastereo- and enantioselectively with an aldehyde or an allylic phosphate to generate valuable multifunctional products. In the second proposed sequence, each point of selectivity, especially the trisubstituted alkene, would be preserved within the final structure. **c**, Sequential catalytic Cu–B(pin) addition/enantioselective allylic substitution, affording products represented by **v**, constitutes an attractive strategy for synthesis of biologically active compounds. NHC, N-heterocyclic carbene; B(pin), (pinacolato)boron; G, functional group.

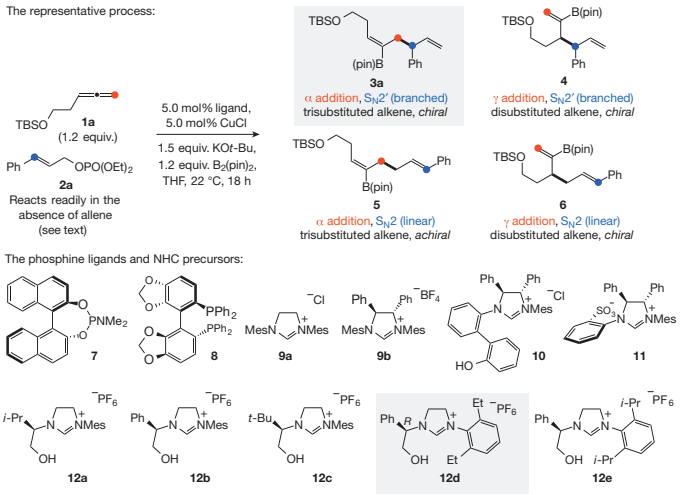
Catalyst identification and method development

Successful implementation of the aforementioned plan demands high chemoselectivity despite the involvement of two C–C π bonds (that is, Cu–B(pin) addition to allene versus allylic phosphate). Monosubstituted allenes[23] as well as allylic carbonates[24] have indeed been shown to undergo efficient reactions with copper–boron complexes. Allenes are comparatively unhindered and might react with a Cu–B(pin) complex more readily, but the Lewis basic phosphate can associate with a transition metal to set off an undesirable sequence of events. Another strategic element is that reaction of the allylcopper intermediate with the allylic phosphate must be followed by a facile reductive elimination (**iv**, Fig. 1b); this way, the trisubstituted alkenylboron unit would be retained and the chiral, branched product isomers would be formed preferentially (that is, **3a** favoured over **4**–**6**; see Table 1).

To identify conditions that would deliver **3a** in favour of 1,1-disubstituted alkenyl–B(pin) **4**, achiral **5** or diene **6** (Table 1), we selected the reaction involving allene **1a** and allylic phosphate **2a**. We soon found that, unlike reactions with aldehydes[6] (Fig. 1b), a phosphine–Cu complex is ineffective (for example, with PCy_3 and **7**[25], entries 1 and 2, Table 1), and bis-phosphine-derived catalysts cause only the allylic phosphate to be consumed (for example, with complex derived from **8**, entry 3). That is, unlike the reactions involving aldehydes, monosubstituted allenes fail to compete with allylic phosphates when bis-phosphines serve as ligands. These observations substantiated our initial concerns regarding the presence of two types of electrophilic olefin.

We then made the unexpected discovery that, in further contrast to carbonyl additions, an NHC–Cu complex can guide the catalytic cycle along the desired path (Table 1). The NHC–Cu species derived from

Table 1 | Examination of copper complexes as catalysts for sequential Cu–B(pin) addition/EAS



Entry number	Ligand or ligand precursor	Conversion (%)*; yield of **3a** (%)†	Site selectivity (**3a**:**4**:**5**:**6**)*	*Z*:*E**	Enantiomeric ratio for **3a**‡
1	PCy_3	<2; NA	NA	NA	NA
2	**7**	<2; NA	NA	NA	NA
3	**8**	>98; <2	NA	NA	NA
4	**9a**	<98; 81	>98:<2:<2:<2	>98:2	NA
5	**9b**	40:<2	NA	NA	NA
6	**10**	>98; 36	>98:<2:<2:<2	>98:2	22:78 (*R*:*S*)
7	**11**	>98; <2	NA	NA	NA
8	**12a**	>98; 67	>98:<2:<2:<2	>98:2	89:11 (*R*:*S*)
9	**12b**	>98; 74	>98:<2:<2:<2	>98:2	93:7 (*R*:*S*)
10	**12c**	>98; 72	>98:<2:<2:<2	>98:2	85:15 (*R*:*S*)
11	**12d**	>98; 80	>98:<2:<2:<2	>98:2	94:6 (*R*:*S*)
12	**12e**	>98; 77	>98:<2:<2:<2	>98:2	92:8 (*R*:*S*)

Reactions were carried out under N_2 atmosphere; see Supplementary Information for details. NA, not applicable; Mes, 2,4,6-Me_3-C_6H_2.
* Conversion, site selectivity and *Z*:*E* ratios were determined by analysis of 400 MHz ^{1}H NMR spectra; variance of values is estimated to be <±2%.
† Yield of purified products.
‡ Enantiomeric ratio values were determined by HPLC analysis; variance of values is estimated to be <±1%. See Supplementary Information for details.

aryl-substituted heterocyclic salt **9a** (entry 4) afforded **3a** as the major component (81% yield); the alternative alkenylboron-containing products **4**, **5** or **6** were not detected (<2%); the complete site (branch) selectivity was equally surprising. Conversely, with enantiomerically pure **9b** (entry 5), **3a** was isolated in trace amounts, and reactions involving chiral salts **10**[26] and **11**[27] either produced **3a** in low yield (compare phenol **10**, entry 6) or none at all (compare sulphonate **11**, entry 7). Investigation of enantiomerically pure NHC precursors bearing an N-aryl and an N-alkyl group led us to establish that reaction with aminoalcohol-derived **12a**[28] (entry 8, Table 1) affords **3a** in 67% yield, >98% S_N2' selectivity and 89:11 e.r. Follow-up studies revealed that enantioselectivity can be sensitive to the substituent at the stereogenic centre of the chiral catalyst (entries 8–10, Table 1). Additional modification revealed that **12d** is precursor to a more efficient (80% yield; entry 11) and enantioselective catalyst (94:6 e.r.). Incorporation of larger N-aryl substituents did not lead to any improvement (compare **12e**, entry 12).

The method can be used to prepare a range of multifunctional organoboron compounds in high selectivity (Fig. 2a). The requisite imidazolinium salt **12d**, an air-stable solid, can be prepared in multigram quantities by a modified version of a reported procedure[28]; the necessary reagents, including either enantiomeric form of phenylglycinol, can be bought at low cost. Allylic phosphates bearing sterically hindered substituents (**3b**, **c**), halogenated aryl groups (**3d**, **e**) or an alkyl unit (**13**) are suitable substrates. Although NHC–Cu–B(pin) complexes react readily with β-alkylstyrenes[29], additions to an allene/EAS occur more readily (**14**). Allenes that contain other modifiable groups, such as an alkyne (**15**), an amine (**16**), or an amide (**17** or **19**) may be used. As the outcomes of the transformations expected to generate amides **17**–**19** indicate, a Lewis basic group, depending on its distance from the allene site, can alter reaction rates. Unsubstituted allene was used to access 1,1-disubstituted alkenyl–B(pin) **20** in 89% yield, >98% branch selectivity and 97:3 e.r. Catalytic cross-coupling reactions with readily accessible aryl halides proceed with retention of stereochemistry to generate trisubstituted alkenes (**21**–**23**, Fig. 2b).

Origins of high efficiency and selectivity

The challenge of designing a multicomponent process is in identifying a catalyst that can clear several efficiency and selectivity hurdles before reaching the finish line and starting again. Key attributes of the optimal NHC–Cu complex (derived from **12d**) are discussed below.

Chemoselectivity and efficiency

The difference between percentage conversion and yield of **3a** with certain Cu complexes (Table 1) signals a breakdown in chemoselectivity: competitive Cu–B addition to **2a** (versus allene **1a**) leads to formation of by-products. Hence, it appears that the less Lewis basic and sterically demanding phosphine-based systems (for example, **8**, Table 1), which are distinct from those of NHC ligands[30,31], allow the phosphate to associate and react more readily. The pathway that hampers the transformations of the less effective NHC–Cu complexes is more tractable. Allylic substitution of a B(pin) group with **2a** produces a branched allylboron intermediate (**24**, Fig. 3a) that can be converted to the corresponding allylcopper species (**25**), which then reacts with another molecule of allylic phosphate **2a** to form 1,5-diene **26**. Indeed, without the allene, the NHC–Cu complexes catalyse the formation of diene **26** efficiently (for example, 53% yield for **9a**, 76% yield for **11** and 50% yield for **12d**). Similar generation of an allylcopper might be inefficient with the Cu centre of a bisphosphine complex, which is probably less Lewis acidic as a result of its weaker two-electron donor ligand (versus an NHC)[32], causing the allylboronate compound to react in other ways.

Comparison of the transformations performed with NHC–Cu complexes derived from **9c**–**f** (Fig. 3b) indicates that the proper balance

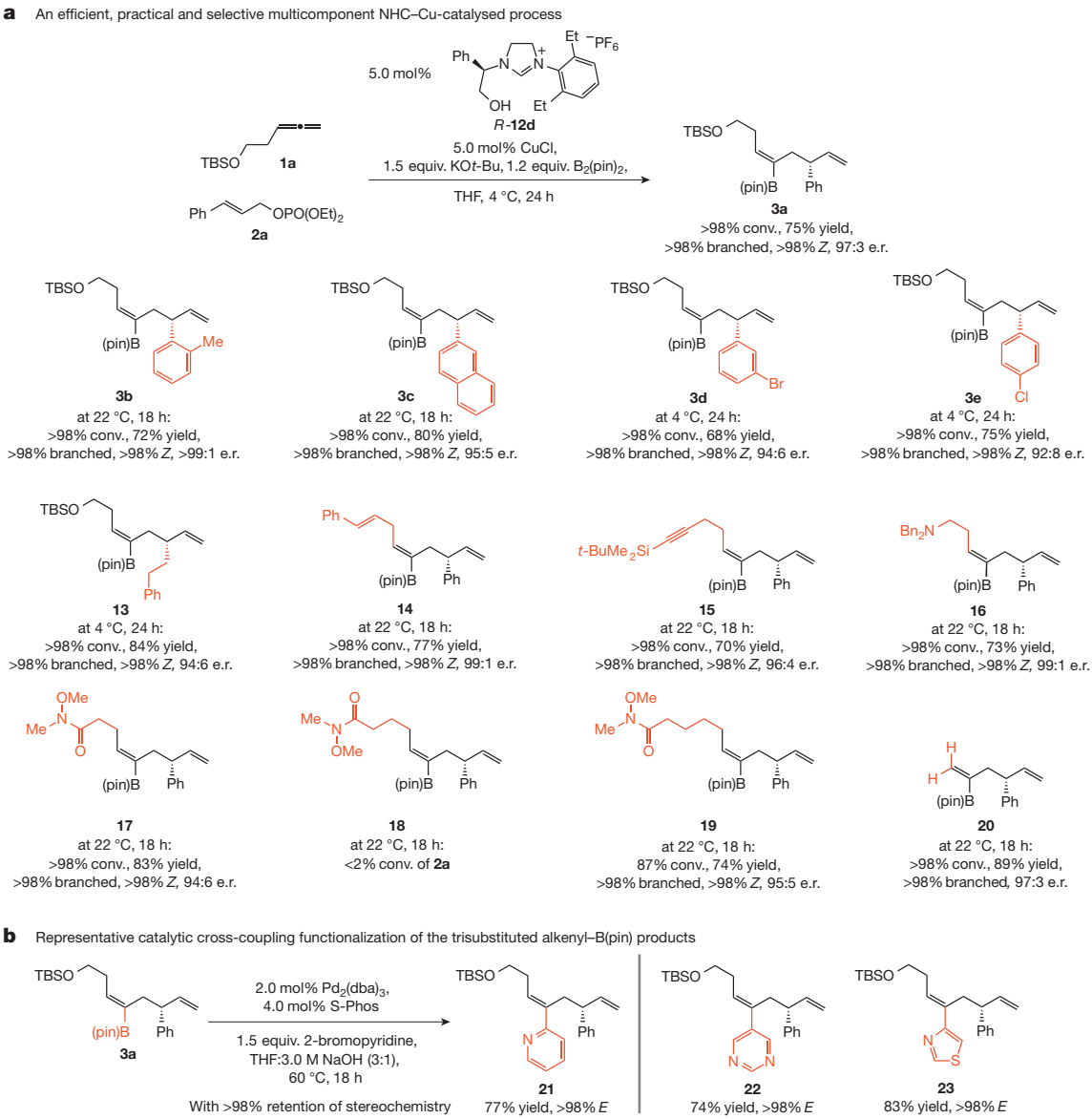


Figure 2 | Catalytic chemo-, site- and enantioselective multicomponent reactions. a, Transformations are promoted by NHC–Cu complexes generated *in situ* from **12d**, which can be easily prepared from inexpensive starting materials on a multigram scale in ~50% overall yield. Transformations proceed with 5.0 mol% catalyst at 4–22 °C and are complete in 18–24 h to deliver the desired products in >98% *Z*, S_N2' and chemoselectivity and 92:8 to >99:1 e.r. **b**, The trisubstituted alkenyl–B(pin) obtained with complete *Z* selectivity can be converted to a variety of trisubstituted *E* alkenes through catalytic cross-coupling with readily available aryl bromides; all reactions proceed with complete retention of stereochemistry. Conv., conversion; TBS, *t*-butyldimethylsilyl; dba, dibenzylideneacetone; S-Phos, 2-dicyclohexylphosphino-2′,6′-dimethoxybiphenyl.

between electronic attributes and size of the heterocyclic ligand is needed if high efficiency and chemoselectivity are to be achieved. The catalyst arising from **9d** is too large to promote transformation, whereas the ligands derived from the smaller **9c** and **9f** contain N-alkyl units (versus N-aryl) and are therefore too nucleophilic to facilitate the desired succession of events. Imidazolinium salt **9e** delivers an NHC–Cu complex that is small enough to promote reaction without being too diminutive or overly nucleophilic to promote Cu–B(pin) addition to the allylic phosphate. The complex resulting from **10** (>98% conversion, 36% yield of **3a**; entry 6, Table 1) in all probability serves as a monodentate ligand that contains an N-mesityl and a smaller *ortho*-substituted N-aryl moiety, rendering it less selective (see below). The bidentate Cu catalyst arising from **11** is probably the only instance of a bidentate complex formed in the screening studies (detailed below); the cuprate species possesses higher-energy Cu *d*-electrons that are more suitable for interacting with the lower-lying π* orbital of an allylic phosphate (versus an allene)[33], facilitating the undesired allylic substitution of a B(pin) unit (low chemoselectivity).

Branch- and enantioselectivity

Chiral heterocyclic ligands (for example, **10**, **11**, **12a–e**) with a chelating group commonly serve as precursors to bidentate NHC–Cu systems (that is, cuprate complexes). However, the resulting Cu–O tether can rupture through reaction with $B_2(pin)_2$, revealing a monodentate complex that carries a neutral metal centre[34]. For two reasons we did not initially think that such a cleavage would take place with Cu complexes resulting from **12a–e**. First, exceptional S_N2' selectivity is usually observed with reactions of organoboron compounds that are promoted by bidentate NHC–Cu catalysts[34,35]; this preference may be attributed to rapid reductive elimination of the Cu(III) intermediate[33] so that substantial steric hindrance can be relieved (compare the top pathway available to **II** in Fig. 3d). The less sterically demanding monodentate complexes, on the other hand, generate achiral linear isomers either preferentially[34] or to a significant degree[35]. Second, high enantioselectivities have been observed with catalysts that contain a chiral NHC ligand that is either bidentate (for example, **11** in Table 1)[28], or monodentate (for example, **9b**) but with conformationally

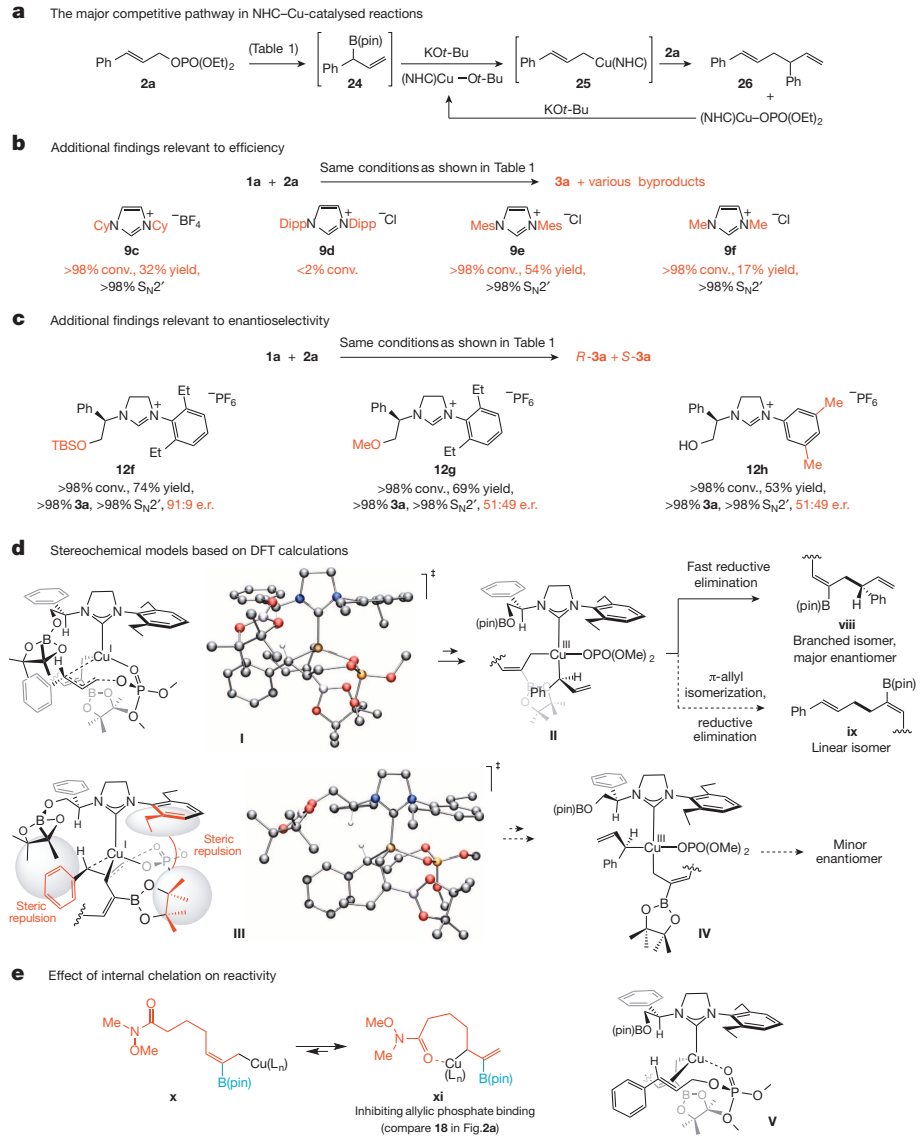


Figure 3 | Origins of high efficiency and selectivity. **a**, Low efficiency with some NHC–Cu-catalysed reactions is due to a competitive pathway arising from undesirable chemoselectivity. **b**, The efficiency of the multicomponent process hinges on the catalyst possessing the appropriate steric and electronic attributes. **c**, Modification of the chiral NHC ligand indicates that the optimal Cu-based catalyst is likely to be a monodentate complex with an N-alkyl side chain containing the sole stereogenic centre. **d**, DFT calculations point to a mode of transformation (**I**) leading to the major enantiomer (versus **III**). The uniformly exceptional branch or S_N2' selectivity (versus linear or S_N2), despite the involvement of a neutral monodentate NHC–Cu catalyst, might be due to steric facilitation of the reductive elimination step (versus π-allyl formation) via **II** and **IV**. **e**, Evidence for the importance of Cu–phosphate chelation to reaction efficiency. Dipp, 2,6-(*i*-Pr)$_2$C$_6$H$_3$; TBS, (*t*-Bu)Me$_2$Si.

constraining stereogenic centres[36,37], or both[24,34,35]. High enantioselectivity without bidentate ligation and/or conformationally restricting substituents is unusual, since it must originate from a single stereogenic centre within the conformationally flexible arm of a C_1-symmetric NHC ligand. Nevertheless, such a scenario became irrefutable when we found that reaction with silyl ether **12f** (Fig. 3c) proceeds with nearly identical efficiency and selectivity as when **12d** is used. (We were unable to prepare and examine an authentic sample of the boronate derivative; similar results were obtained with the corresponding *tert*-butyldiphenylsilyl ether analogue of **12f**.) Additionally, with methyl ether **12g** (Fig. 3c), enantioselectivity was all but completely eroded. Stereoselectivity is thus likely to be induced by the large B(pin)-substituted chiral appendage, formed through reaction of $B_2(pin)_2$ with the Cu–O bond and emulated by the silyl ether in **12f**.

Calculations through the use of density functional theory (DFT) point to transition structure **I** as the source of the major product enantiomer (Fig. 3d; see Supplementary Information for details of all calculations). The allylic phosphate occupies two sites of the tetrahedral Cu(I) complex to generate a square planar Cu(III) species that undergoes reductive elimination via **II** to give **viii** (versus **ix**). The P=O→Cu coordination facilitates the association of the allylic phosphate with the sterically demanding NHC–Cu–allyl complex; this picture is supported by the variations in reaction efficiency observed for the transformations involving products **17**–**19** (Fig. 2a). In the case of **18** (<2% conversion), the Lewis basic amide carbonyl is properly situated to chelate with the Cu centre in the first intermediate to prevent phosphate chelation (**x**, Fig. 3e). The ring size in the bidentate complex **xi** is similar to that found in the oxidative addition precursor **V** (Fig. 3e).

The two-point catalyst/substrate binding enhances the organization of the stereochemistry-determining transition state, generating high stereochemical induction via **II**. The minor isomer is probably produced via **III**, wherein the sizeable boronate group can swerve into close contact with the protruding allylic phosphate substituent. The B(pin) moiety of the allyl ligand must either collide with the ethyl substituents of the

NHC's N-aryl moiety (shown) or induce steric repulsion due to proximity of the B(pin) unit and the NHC side chain. There must therefore be a feature of the catalyst structure that is responsible for C–C bond formation occurring near the chiral arm of the NHC ligand. Molecular models suggest that, because of steric factors, the ortho (ethyl) substituents of the ligand's N-aryl moiety discourage placement of the allyl fragments in their vicinity. The complete loss of enantioselectivity that results from placement of the groups at the N-aryl moiety's C3 and C5 positions corroborates the proposed scenario (that is, the derived boronate of NHC precursor **12h**, Fig. 3c).

Exceptional $S_N2':S_N2$ ratios

Then there are the exceptional $S_N2':S_N2$ ratios despite involvement of a monodentate–Cu complex[35,38]. This almost certainly originates from the sizeable B(pin) unit of the allyl ligand of the Cu(III) complex; the boronate moiety is absent in the formerly examined EAS reactions with organoboron compounds[34,35]. The steric repulsion engendered by the sizeable B(pin) group elevates the ground state energy of the Cu(III) intermediate species **II** (major) and **IV** (minor), accelerating reductive elimination (to give **viii** in Fig. 3d) before it can collapse to the π-allyl species (**ix** in Fig. 3d). DFT calculations support the contention that the presence of the large B(pin) group lowers the activation barrier to reductive elimination (strain release).

Gram-scale total synthesis of rottnestol

Synthesis of a complex organic molecule with catalytic multicomponent processes as its central feature would be a clear indicator of the utility of such processes, particularly if meaningful quantities of a target molecule were to be secured. We first designed a route to prepare gram quantities of pure rottnestol where every issue of stereochemical control would be addressed by a catalytic transformation. We envisioned using the NHC–Cu-catalysed enantioselective process involving an allylic phosphate for synthesis of the polyene segment, while the carbohydrate moiety would be accessed through a catalytic $B_2(pin)_2$/allene/aldehyde fusion (Fig. 4).

The synthesis route commenced with the Cu–B(pin) addition/EAS sequence (Fig. 4a). Treatment of monosubstituted allene **1b** and methyl-substituted allylic phosphate **2b** with 3.0 mol% *S*-**12d** and CuCl (Fig. 2) afforded trisubstituted alkenyl–B(pin) **27** in 79% yield, with complete branch and *Z* selectivity and in 92:8 e.r.; the reaction was performed on two batches of ~1.7 g of **1b**, delivering a total of ~4.2 g of the product. Conversion of the C–B(pin) to a C–Me bond was accomplished with complete inversion of stereochemistry (>98% *E*) through reaction with methyllithium and iodine[39], delivering trisubstituted alkene **28** in 91% yield (3.4 g). NHC–Ru-catalysed *E*-selective cross-metathesis with commercially available vinyl–B(pin)[16] in the presence of 5.0 mol% Ru carbene **29**[40] and formation of the corresponding alkenyl–iodide[16] proceeded in 80% overall yield, furnishing ~3.6 g of **30**, which was transformed to 1.4 g of triene **31** in three straightforward steps (73% overall yield).

To prepare the carbohydrate fragment (Fig. 4b), we adopted an enantioselective reaction involving aldehyde **32**, which can be prepared in one step from a commercially available alcohol and four other entities that can also be purchased: methylallene **1c**, $B_2(pin)_2$, bis-phosphine **33** and CuCl. Silyl protection of the β-hydroxyketone afforded **34** in 75% overall yield (3.4 g through two batches) with complete control of

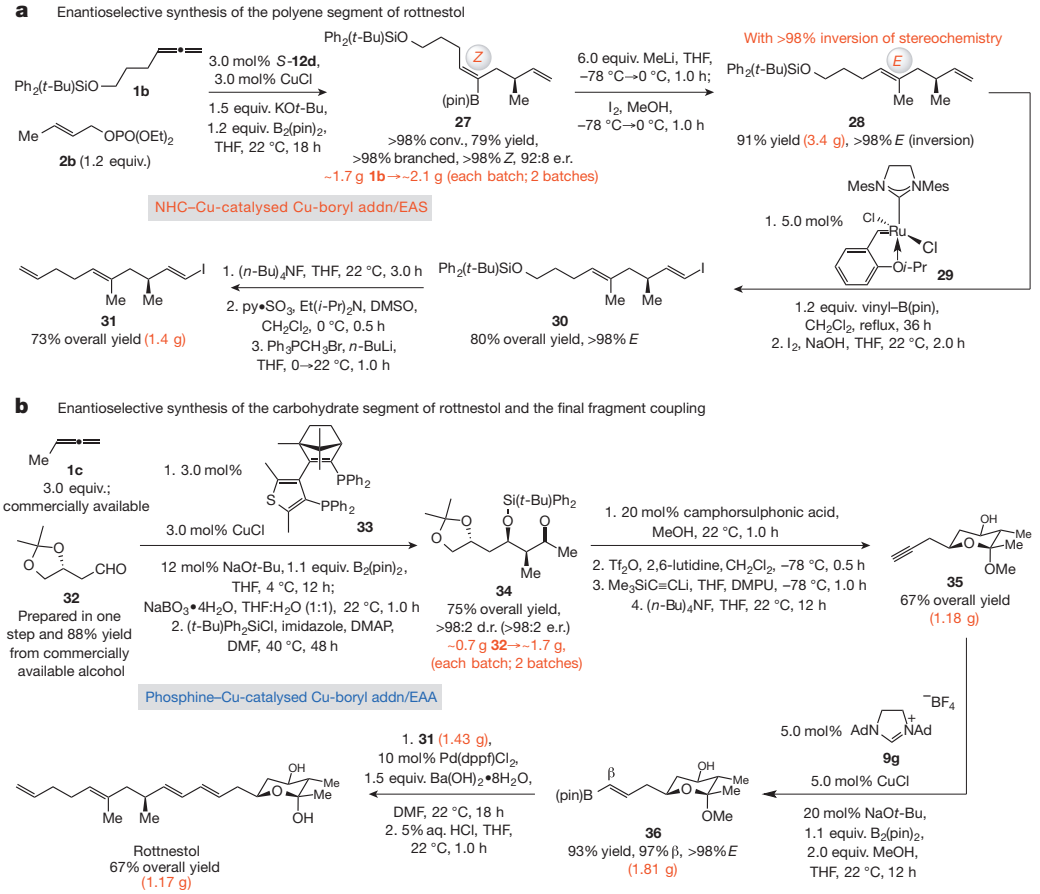


Figure 4 | Enantioselective gram-scale synthesis of rottnestol. Every stereochemical issue in the route is addressed by a catalytic process that involves an organoboron compound; this is highlighted by two multicomponent chemo-, site-, diastereo- and enantioselective assemblies. **a**, Site- and enantioselective NHC–Cu-catalysed $B_2(pin)_2$/allene/allylic phosphate and NHC–Ru-catalysed catalytic cross-metathesis (CM) reactions are combined to access the acyclic fragment. **b**, A phosphine–Cu-catalysed multicomponent process involving an allene and an aldehyde is used to access the carbohydrate moiety. The final fragment coupling is achieved by phosphine–Pd-catalysed coupling, generating nearly 1.2 g of the natural product. DMAP, 4-dimethylaminopyridine; DMPU, N,N′-dimethyl-N,N′-trimethyleneurea; dppf, 1,1′-bis(diphenylphosphino) ferrocene. Ad, adamantyl.

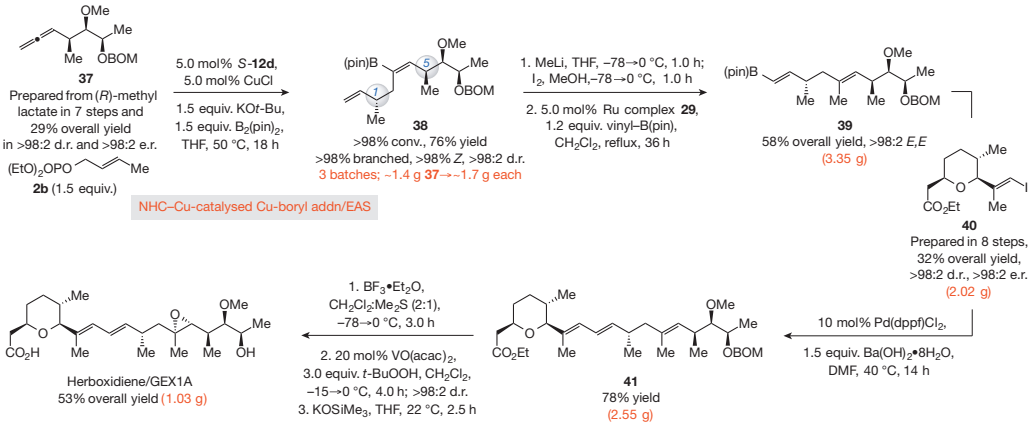


Figure 5 | Enantioselective gram-scale synthesis of herboxidiene. The key step, taking place at mid-point in the multistep route, involves a relatively complex enantiomerically pure monosubstituted allene (**37**). The reaction can be performed on gram-scale batches to obtain ~1.7 g of 1,5-diene **38** for each run (~76% yield), with >98% site-, *Z*- and diastereoselectivity. Subsequent conversion to *E*,*E*-diene **39** proceeds with complete stereochemical control as well. Catalytic cross-coupling generates triene **41**, which is then transformed to 1.03 g of the natural product. It is noteworthy that every transformation shown above that involves a stereochemical issue proceeds with complete selectivity (that is, catalytic multicomponent process, alkylation of the alkenylboron intermediate, catalytic cross-metathesis, catalytic cross-coupling and directed epoxidation). acac, acetylacetone.

stereoselectivity (>98:2 d.r. and e.r.). Ketone **34** was converted to carbohydrate **35** after four steps in 67% overall yield (1.18 g). NHC–Cu-catalysed protoboration of the terminal alkyne in **35** furnished β-alkenyl–B(pin) **36** in 97:3 β:α ratio, >98% *E* selectivity and 93% yield (1.8 g)[41]. More than one gram of stereoisomerically pure rottnestol was obtained after catalytic cross-coupling[42] of alkenyl–iodide **31** and alkenyl–B(pin) **36** followed by generation of the cyclic hemiacetal by acid treatment. The route in Fig. 4 is more efficient than those reported previously (21.5% versus 3.7% overall yield)[43] and which resulted in no more than milligram quantities of the target molecule.

Gram-scale total synthesis of herboxidiene

Devising a route leading to anti-tumour agent herboxidiene was next. Here, we explored a different aspect of the NHC–Cu-catalysed transformation (Fig. 5). In the case of rottnestol, the multicomponent process was employed early on (**27**, Fig. 4); in contrast, with herboxidiene, the process would be implemented at a later stage with a more complex allene. In the event, ~7 g of substrate **37** were obtained by a seven-step procedure in 29% overall yield, >98:2 d.r. and e.r.; the central reaction in the sequence was phosphine–Cu-catalysed multicomponent reaction of $B_2(pin)_2$, methylallene **1c** and an aldehyde derived from (*R*)-methyl lactate (compare synthesis of **34**). Considerable structural complexity, including the appropriate 1,5-relative stereochemistry, was thus generated in short order: 1,5-diene **38** (~1.7 g for each run) was obtained in ~76% yield with complete site-, *Z*- and diastereoselectivity. Trisubstituted olefin **39** was accessed through alkylation and catalytic cross-metathesis with vinyl–B(pin), yielding ~3.3 g of the desired product; both alkenes were formed with >98% *E* selectivity. Phosphine–Pd-catalysed cross-coupling of alkenyl–B(pin) **39** with alkenyl-iodide **40**, synthesized through a diastereo- and enantioselective eight-step process starting from β-(+)-citronellene (see Supplementary Information), afforded 2.55 g of triene **41**. After three operations[44,45], 1.03 g of the anti-tumour agent was secured; this represents an overall yield nearly twice that of the most concise of the previously reported syntheses[45] (5.5% versus 3.4%; see Supplementary Information for bibliography).

Conclusions and discussions

The advances outlined here demonstrate that two simple unsaturated organic molecules and a commercially available diboron reagent can be combined to generate multifunctional alkenylboron fragments that are marked by several advantageous attributes. The requisite catalyst is assembled *in situ* by the reaction of abundant and inexpensive CuCl and a chiral ligand that is synthesized in multigram quantities readily and cost-effectively. Owing to the above features, and because the NHC–Cu-catalysed process is robust, gram quantities of a variety of complex organic molecules become reliably available.

This advance foreshadows the development of protocols involving additional difficult-to-access alkenylboron-containing organocopper compounds. What emerges is the possibility of using other abundantly available poly-unsaturated hydrocarbons, such as dienes[46] or enynes[47,48], for efficient preparation of high-value products. Such a strategy obviates the need for succumbing to one-at-a-time installation of each functional unit, resulting in pathways that are unnecessarily time consuming, costly and waste generating.

Received 2 July; accepted 5 August 2014.

1. Ramón, D. J. & Yus, M. Asymmetric multicomponent reactions (AMCRs): the new frontier. *Angew. Chem. Int. Edn* **44,** 1602–1634 (2005).
2. Ruijter, E., Scheffelaar, R. & Orru, R. V. A. Multicomponent reaction design in the quest for molecular complexity and diversity. *Angew. Chem. Int. Edn* **50,** 6234–6246 (2011).
3. Bower, J. F., Kim, I. S., Patman, R. L. & Krische, M. J. Catalytic carbonyl addition through transfer hydrogenation: a departure from preformed organometallic reagents. *Angew. Chem. Int. Edn* **48,** 34–46 (2009).
4. Ngai, M.-Y., Barchuk, A. & Krische, M. J. Enantioselective iridium-catalyzed imine vinylation. Optically enriched allylic amines via alkyne–imine reductive coupling mediated by hydrogen. *J. Am. Chem. Soc.* **129,** 12644–12645 (2007).
5. Hassan, A. & Krische, M. J. Unlocking hydrogenation for C–C bond formation: a brief overview of enantioselective methods. *Org. Process Res. Dev.* **15,** 1236–1242 (2011).
6. Meng, F., Jang, H., Jung, B. & Hoveyda, A. H. Cu-catalyzed chemoselective preparation of 2-(pinacolato)boron-substituted allylcopper complexes and their in situ site-, diastereo-, and enantioselective additions to aldehydes and ketones. *Angew. Chem. Int. Edn* **52,** 5046–5051 (2013).
7. Takano, D. *et al.* Total synthesis of nafuredin, a selective NADH-fumarate reductase inhibitor. *Org. Lett.* **3,** 2289–2291 (2001).
8. Schow, S. R., Bloom, J. D., Thompson, A. S., Winzenberg, K. N. & Smith, A. B. Milbemycin–avermectin studies. 5. Total synthesis of milbemycin β_3 and its C(12) epimer. *J. Am. Chem. Soc.* **108,** 2662–2674 (1986).
9. Zhang, P., Le, H., Kyne, R. E. & Morken, J. P. Enantioselective construction of all-carbon quaternary centers by branch-selective Pd-catalyzed allyl–allyl cross-coupling. *J. Am. Chem. Soc.* **133,** 9716–9719 (2011).
10. Hornillos, V., Peréz, M., Fañanás-Mastral, M. & Feringa, B. L. Copper-catalyzed enantioselective allyl–allyl cross-coupling. *J. Am. Chem. Soc.* **135,** 2140–2143 (2013).
11. Hamilton, J. Y., Sarlah, D. & Carreira, E. M. Iridium-catalyzed enantioselective allyl–alkene coupling. *J. Am. Chem. Soc.* **136,** 3006–3009 (2014).
12. Tanaka, H. *et al.* Structure of FK506: a novel immunosuppressant isolated from *Streptomyces*. *J. Am. Chem. Soc.* **109,** 5031–5033 (1987).
13. Jones, T. K., Reamer, R. A., Desmond, R. & Mills, S. G. Chemistry of tricarbonyl hemiketals and application of Evans' technology to the total synthesis of the immunosuppressant (–)-FK-506. *J. Am. Chem. Soc.* **112,** 2998–3017 (1990).
14. Nakatsuka, M. *et al.* Total synthesis of FK-506 and an FKBP probe reagent, (C_8,C_9-$^{13}C_2$)-FK-506. *J. Am. Chem. Soc.* **112,** 5583–5601 (1990).

15. Ireland, R. E., Gleason, J. L., Gegnas, L. D. & Highsmith, T. K. A total synthesis of FK-506. *J. Org. Chem.* **61,** 6856–6872 (1996).
16. Morrill, C. & Grubbs, R. H. Synthesis of functionalized vinyl boronates via ruthenium-catalyzed olefin cross-metathesis and subsequent conversion to vinyl halides. *J. Org. Chem.* **68,** 6031–6034 (2003).
17. Kotha, S., Lahiri, K. & Kashinath, D. Recent applications of the Suzuki-Miyaura cross-coupling reaction in organic synthesis. *Tetrahedron* **58,** 9633–9695 (2002).
18. Takano, D. *et al.* Absolute configuration of nafuredin, a new specific NADH-fumarate reductase inhibitor. *Tetrahedr. Lett.* **42,** 3017–3020 (2001).
19. Omura, S. *et al.* An anthelmintic compound, nafuredin, shows selective inhibition of complex I in helminth mitochondria. *Proc. Natl Acad. Sci. USA* **98,** 60–62 (2001).
20. Shoop, W. L., Mrozik, H. & Fisher, M. H. Structure and activity of avermectins and milbemycins in animal health. *Vet. Parasitol.* **59,** 139–156 (1995).
21. Erickson, K. L., Beutler, J. A., Cardellina, J. H. & Boyd, M. R. Rottnestol, a new hemiketal from the sponge *Haliclona* sp. *Tetrahedron* **51,** 11953–11958 (1995).
22. Hasegawa, M. *et al.* Identification of SAP155 as the target of GEX1A (Herboxidiene), an antitumor natural product. *ACS Chem. Biol.* **6,** 229–233 (2011).
23. Meng, F., Jung, B., Haeffner, F. & Hoveyda, A. H. NHC–Cu-catalyzed protoboration of monosubstituted allenes. Ligand-controlled site selectivity, application to synthesis and mechanism. *Org. Lett.* **15,** 1414–1417 (2013).
24. Guzman-Martinez, A. & Hoveyda, A. H. Enantioselective synthesis of allylboronates bearing a tertiary or quaternary B-substituted stereogenic carbon by NHC–Cu-catalyzed substitution reactions. *J. Am. Chem. Soc.* **132,** 10634–10637 (2010).
25. De Vries, A. H. M., Meetsma, A. & Feringa, B. L. Enantioselective conjugate addition of dialkylzinc reagents to cyclic and acyclic enones catalyzed by chiral copper complexes of new phosphorus amidites. *Angew. Chem. Int. Edn* **35,** 2374–2376 (1996).
26. Van Veldhuizen, J. J., Campbell, J. E., Giudici, R. E. & Hoveyda, A. H. A readily available chiral Ag-based N-heterocyclic carbene complex for use in efficient and highly enantioselective Ru-catalyzed olefin metathesis and Cu-catalyzed allylic alkylation reaction. *J. Am. Chem. Soc.* **127,** 6877–6882 (2005).
27. May, T. L., Brown, M. K. & Hoveyda, A. H. Enantioselective synthesis of all-carbon quaternary stereogenic centers by catalytic asymmetric conjugate additions of alkyl and aryl aluminum reagents to five-, six-, and seven-membered-ring β-substituted cyclic enones. *Angew. Chem. Int. Edn* **47,** 7358–7362 (2008).
28. Clavier, H., Coutable, L., Toupet, L., Guillemin, J.-C. & Mauduit, M. Design and synthesis of new bidentate alkoxy–NHC ligands for enantioselective copper-catalyzed conjugate addition. *J. Organomet. Chem.* **690,** 5237–5254 (2005).
29. Lee, Y. & Hoveyda, A. H. Efficient boron–copper additions to aryl-substituted alkenes promoted by NHC-based catalysts. Enantioselective Cu-catalyzed hydroboration reactions. *J. Am. Chem. Soc.* **131,** 3160–3161 (2009).
30. Díez-González, S. & Nolan, S. P. Stereoelectronic parameters associated with N-heterocyclic carbene (NHC) ligands: a quest for understanding. *Coord. Chem. Rev.* **251,** 874–883 (2007).
31. Maji, B., Breugst, M. & Mayr, H. N-Heterocyclic carbenes: organocatalysts with moderate nucleophilicity but extraordinarily high Lewis basicity. *Angew. Chem. Int. Edn* **50,** 6915–6919 (2011).
32. Denmark, S. E. & Beutner, G. L. Lewis base catalysis in organic synthesis. *Angew. Chem. Int. Edn* **47,** 1560–1638 (2008).
33. Yoshikai, N. & Nakamura, E. Mechanisms of nucleophilic organocopper(I) reactions. *Chem. Rev.* **112,** 2339–2372 (2012).
34. Jung, B. & Hoveyda, A. H. Site- and enantioselective formation of allene-bearing tertiary or quaternary carbon stereogenic centers through NHC–Cu-catalyzed allylic substitution. *J. Am. Chem. Soc.* **134,** 1490–1493 (2012).
35. Gao, F., Carr, J. L. & Hoveyda, A. H. A broadly applicable NHC–Cu-catalyzed approach for efficient, site-, and enantioselective coupling of readily accessible (pinacolato)alkenylboron compounds to allylic phosphates and applications to natural product synthesis. *J. Am. Chem. Soc.* **136,** 2149–2161 (2014).
36. Park, J. K., Lackey, H. H., Ondrusek, B. A. & McQuade, D. T. Stereoconvergent synthesis of chiral allylboronates from an *E/Z* mixture of allylic aryl ethers using 6-NHC–Cu(I) catalyst. *J. Am. Chem. Soc.* **133,** 2410–2413 (2011).
37. Lee, K.-s. & Hoveyda, A. H. Monodentate Non-C_2-symmetric chiral *N*-heterocyclic carbene complexes for enantioselective synthesis. Cu-catalyzed conjugate additions of aryl- and alkenylsilylfluorides to cyclic enones. *J. Org. Chem.* **74,** 4455–4462 (2009).
38. Gao, F., Lee, Y., Mandai, K. & Hoveyda, A. H. Quaternary carbon stereogenic centers through copper-catalyzed enantioselective allylic substitutions with readily accessible aryl- or heteroaryllithium reagents and aluminum chlorides. *Angew. Chem. Int. Edn* **49,** 8370–8374 (2010).
39. Xu, S., Lee, C.-T., Rao, H. & Negishi, E. Highly (≥98%) stereo- and regioselective trisubstituted alkene synthesis of wide applicability via 1-halo-1-alkyne hydroboration-tandem Negishi–Suzuki coupling or organoborate migratory insertion. *Adv. Synth. Catal.* **353,** 2981–2987 (2011).
40. Garber, S. B., Kingsbury, J. S., Gray, B. L. & Hoveyda, A. H. Efficient and recyclable monomeric and dendritic Ru-based metathesis catalysts. *J. Am. Chem. Soc.* **122,** 8168–8179 (2000).
41. Jang, H., Zhugralin, A. R., Lee, Y. & Hoveyda, A. H. Highly selective methods for synthesis of internal (α-) vinylboronates through efficient NHC–Cu-catalyzed hydroboration of terminal alkynes. Utility in chemical synthesis and mechanistic basis for selectivity. *J. Am. Chem. Soc.* **133,** 7859–7871 (2011).
42. Fürstner, A. *et al.* Total synthesis of lejimalide A–D and assessment of the remarkable actin-depolymerizing capacity of these polyene macrolides. *J. Am. Chem. Soc.* **129,** 9150–9161 (2007).
43. Czuba, I. R., Zammit, S. & Rizzacasa, M. A. Total synthesis of marine sponge metabolites (+)-rottnestol, (+)-raspailol A and (+)-raspailol B. *Org. Biomol. Chem.* **1,** 2044–2056 (2003).
44. Pellicena, M., Krämer, K., Romea, P. & Urpí, F. Total synthesis of (+)-herboxidiene from two chiral lactate-derived ketones. *Org. Lett.* **13,** 5350–5353 (2011).
45. Murray, T. J. & Forsyth, C. J. Total synthesis of GEX1A. *Org. Lett.* **10,** 3429–3431 (2008).
46. Sasaki, Y., Zhong, C., Sawamura, M. & Ito, H. Copper(I)-catalyzed asymmetric monoborylation of 1,3-dienes: synthesis of enantioenriched cyclic homoallyl- and allylboronates. *J. Am. Chem. Soc.* **132,** 1226–1227 (2010).
47. Sasaki, Y., Horita, Y., Zhong, C., Sawamura, M. & Ito, H. Copper(I)-catalyzed regioselective monoborylation of 1,3-enynes with an internal triple bond: selective synthesis of 1,3-dienylboronates and 3-alkynylboronates. *Angew. Chem. Int. Ed.* **50,** 2778–2782 (2011).
48. Meng, F., Haeffner, J. & Hoveyda, A. H. Diastereo- and enantioselective reactions of bis(pinacolato)diboron, 1,3-enynes, and aldehydes catalyzed by an easily accessible bisphosphine–Cu complex. *J. Am. Chem. Soc.* **136,** 11304–11307 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements This research was supported by grants from the National Institutes of Health, Institute of General Medical Sciences (GM-47480) and the National Science Foundation (CHE-1111074 and CHE-1362763). F.M. acknowledges a LaMattina graduate fellowship in organic synthesis. We thank M. J. Koh, D. L. Silverio and F. Haeffner for discussions, Boston College for access to computational facilities and Frontier Scientific, Inc., for gifts of $B_2(pin)_2$.

Author Contributions F.M. performed the catalyst studies and method development studies, as well as the total syntheses of rottnestol and herboxidiene. K.P.M. carried out the computational studies. A.H.H. and F.M. conceived the project. A.H.H. designed and directed the investigations and composed the manuscript with revisions provided by the other authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.H.H. (amir.hoveyda@bc.edu).

ARTICLE

OPEN

doi:10.1038/nature13726

The genomic substrate for adaptive radiation in African cichlid fish

David Brawand[1,2]*, Catherine E. Wagner[3,4]*, Yang I. Li[2]*, Milan Malinsky[5,6], Irene Keller[4], Shaohua Fan[7], Oleg Simakov[7,8], Alvin Y. Ng[9], Zhi Wei Lim[9], Etienne Bezault[10], Jason Turner-Maier[1], Jeremy Johnson[1], Rosa Alcazar[11], Hyun Ji Noh[1], Pamela Russell[12], Bronwen Aken[6], Jessica Alföldi[1], Chris Amemiya[13], Naoual Azzouzi[14], Jean-François Baroiller[15], Frederique Barloy-Hubler[14], Aaron Berlin[1], Ryan Bloomquist[16], Karen L. Carleton[17], Matthew A. Conte[17], Helena D'Cotta[15], Orly Eshel[18], Leslie Gaffney[1], Francis Galibert[14], Hugo F. Gante[19], Sante Gnerre[1], Lucie Greuter[3,4], Richard Guyon[14], Natalie S. Haddad[16], Wilfried Haerty[2], Rayna M. Harris[20], Hans A. Hofmann[20], Thibaut Hourlier[6], Gideon Hulata[18], David B. Jaffe[1], Marcia Lara[1], Alison P. Lee[9], Iain MacCallum[1], Salome Mwaiko[3], Masato Nikaido[21], Hidenori Nishihara[21], Catherine Ozouf-Costaz[22], David J. Penman[23], Dariusz Przybylski[1], Michaelle Rakotomanga[14], Suzy C. P. Renn[10], Filipe J. Ribeiro[1], Micha Ron[18], Walter Salzburger[19], Luis Sanchez-Pulido[2], M. Emilia Santos[19], Steve Searle[6], Ted Sharpe[1], Ross Swofford[1], Frederick J. Tan[24], Louise Williams[1], Sarah Young[1], Shuangye Yin[1], Norihiro Okada[21,25], Thomas D. Kocher[17], Eric A. Miska[5], Eric S. Lander[1], Byrappa Venkatesh[9], Russell D. Fernald[11], Axel Meyer[7], Chris P. Ponting[2], J. Todd Streelman[16], Kerstin Lindblad-Toh[1,26], Ole Seehausen[3,4] & Federica Di Palma[1,27]

Cichlid fishes are famous for large, diverse and replicated adaptive radiations in the Great Lakes of East Africa. To understand the molecular mechanisms underlying cichlid phenotypic diversity, we sequenced the genomes and transcriptomes of five lineages of African cichlids: the Nile tilapia (*Oreochromis niloticus*), an ancestral lineage with low diversity; and four members of the East African lineage: *Neolamprologus brichardi/pulcher* (older radiation, Lake Tanganyika), *Metriaclima zebra* (recent radiation, Lake Malawi), *Pundamilia nyererei* (very recent radiation, Lake Victoria), and *Astatotilapia burtoni* (riverine species around Lake Tanganyika). We found an excess of gene duplications in the East African lineage compared to tilapia and other teleosts, an abundance of non-coding element divergence, accelerated coding sequence evolution, expression divergence associated with transposable element insertions, and regulation by novel microRNAs. In addition, we analysed sequence data from sixty individuals representing six closely related species from Lake Victoria, and show genome-wide diversifying selection on coding and regulatory variants, some of which were recruited from ancient polymorphisms. We conclude that a number of molecular mechanisms shaped East African cichlid genomes, and that amassing of standing variation during periods of relaxed purifying selection may have been important in facilitating subsequent evolutionary diversification.

Wide variation in the rates of diversification among lineages is a feature of evolution that has fascinated biologists since Darwin[1,2]. With approximately 2,000 known species, hundreds of which coexist in individual African lakes, cichlid fish are amongst the most striking examples of adaptive radiation, the phenomenon whereby a single lineage diversifies into many ecologically varied species in a short span of time[3] (Fig. 1). The largest radiations, which in Lakes Victoria, Malawi and Tanganyika, have generated between 250 (Tanganyika) and 500 (Malawi and Victoria) species per lake, took no more than 15,000 to 100,000 years for Victoria and less than 5 million years for Malawi[3–5], but 10–12 million years for Lake Tanganyika[6]. The radiations in Lake Victoria and Malawi thus display the highest sustained rates of speciation known to date in vertebrates[7]. The evolution of these lineages and their genomes has presumably been shaped by cycles of population expansion, fragmentation and contraction as lineages colonized lakes, diversified, collapsed when lakes dried up, and re-colonized lakes, and by episodic adaptation to a multitude of ecological niches coupled with strong sexual selection. Genetic diversity within lake radiations has been influenced by admixture following multiple colonization events and periodic infusions through hybridization[8,9].

Cichlid phenotypic diversity encompasses variation in behaviour, body shape, coloration and ecological specialization. The frequent occurrence of convergent evolution of similar ecotypes (Fig. 1) suggests a primary role of natural selection in shaping cichlid phenotypic diversity[10,11]. In addition, the importance of sexual selection is demonstrated by a profusion of exaggerated sexually dimorphic traits like male nuptial colour and elaborate bower building by males[3]. Ecological and sexual selection

[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. [2]MRC Functional Genomics Unit, University of Oxford, Oxford OX1 3QX, UK. [3]Department of Fish Ecology and Evolution, Eawag Swiss Federal Institute of Aquatic Science and Technology, Center for Ecology, Evolution & Biogeochemistry, CH-6047 Kastanienbaum, Switzerland. [4]Division of Aquatic Ecology, Institute of Ecology & Evolution, University of Bern, CH-3012 Bern, Switzerland. [5]Gurdon Institute, Cambridge CB2 1QN, UK. [6]Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK. [7]Department of Biology, University of Konstanz, D-78457 Konstanz, Germany. [8]European Molecular Biology Laboratory, 69117 Heidelberg, Germany. [9]Institute of Molecular and Cell Biology, A*STAR, 138673 Singapore. [10]Department of Biology, Reed College, Portland, Oregon 97202, USA. [11]Biology Department, Stanford University, Stanford, California 94305-5020, USA. [12]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California 91125, USA. [13]Benaroya Research Institute at Virginia Mason, Seattle, Washington 98101, USA. [14]Institut Génétique et Développement, CNRS/University of Rennes, 35043 Rennes, France. [15]CIRAD, Campus International de Baillarguet, TA B-110/A, 34398 Montpellier cedex 5, France. [16]School of Biology, Georgia Institute of Technology, Atlanta, Georgia 30332-0230, USA. [17]Department of Biology, University of Maryland, College Park, Maryland 20742, USA. [18]Animal Genetics, Institute of Animal Science, ARO, The Volcani Center, Bet-Dagan, 50250 Israel. [19]Zoological Institute, University of Basel, CH-4051 Basel, Switzerland. [20]Department of Integrative Biology, Center for Computational Biology and Bioinformatics; The University of Texas at Austin, Austin, Texas 78712, USA. [21]Department of Biological Sciences, Tokyo Institute of Technology, Tokyo, 226-8501 Yokohama, Japan. [22]Systématique, Adaptation, Evolution, National Museum of Natural History, 75005 Paris, France. [23]Institute of Aquaculture, University of Stirling, Stirling FK9 4LA, UK. [24]Carnegie Institution of Washington, Department of Embryology, 3520 San Martin Drive Baltimore, Maryland 21218, USA. [25]National Cheng Kung University, Tainan City, 704 Taiwan. [26]Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, 751 23 Uppsala, Sweden. [27]Vertebrate and Health Genomics, The Genome Analysis Centre, Norwich NR18 7UH, UK.

*These authors contributed equally to this work.

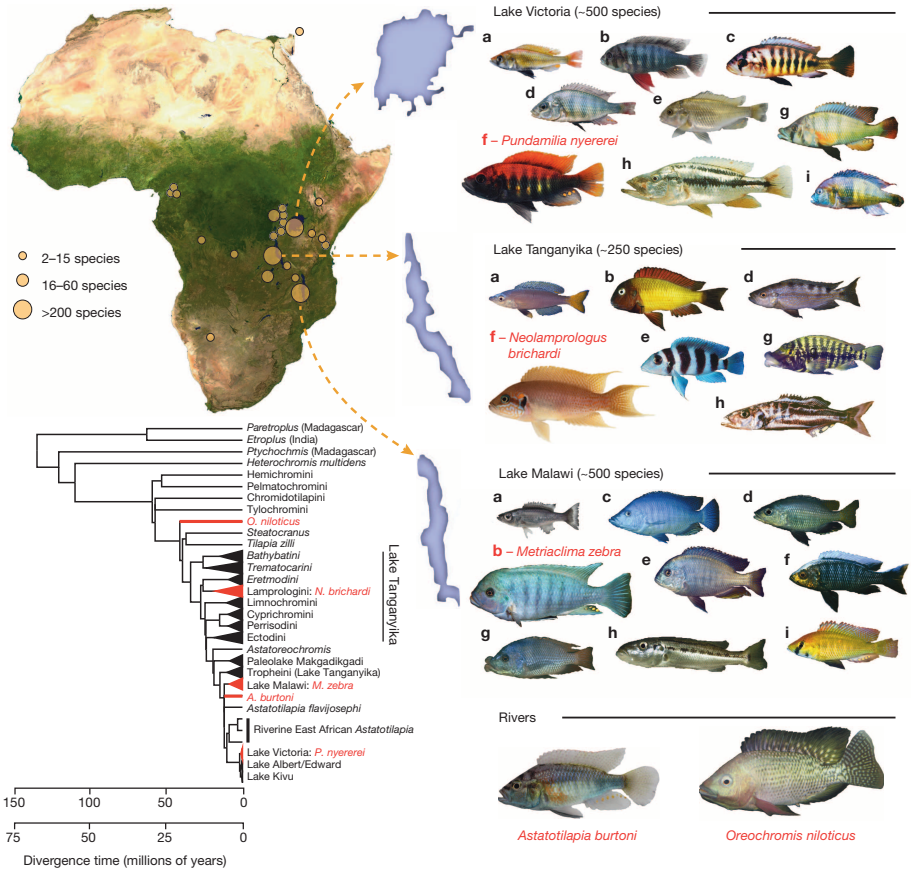


Figure 1 | The adaptive radiation of African cichlid fish. Top left, map of Africa showing lakes in which cichlid fish have radiated. Right, the five sequenced species: *Pundamilia nyererei* (endemic of Lake Victoria); *Neolamprologus brichardi* (endemic of Lake Tanganyika); *Metriaclima zebra* (endemic of Lake Malawi); *Oreochromis niloticus* (from rivers across northern Africa); *Astatotilapia burtoni* (from rivers connected to Lake Tanganyika). Major ecotypes are shown from each lake: **a**, pelagic zooplanktivore; **b**, rock-dwelling algae scraper; **c**, paedophage (absent from Lake Tanganyika); **d**, scale eater; **e**, snail crusher; **f**, reef-dwelling planktivore; **g**, lobe-lipped insect eater; **h**, pelagic piscivore; **i**, ancestral river-dweller also found in lakes (absent from Lake Tanganyika). Bottom left, phylogenetic tree illustrating relationships between the five sequenced species (red), major adaptive radiations and major river lineages. The tree is from ref. 4, pruned to the major lineages. Upper timescale (4), lower timescale (32). Photos by Ad Konings (Tanganyika **a**, **b**, **d**, **e**, **g**, **h**; Malawi **a**, **c**, **d**, **e**, **f**, **g**, **h**, **i**), O.S. (Victoria **a**–**g**, **i**; Malawi **b**), Frans Witte (Victoria **h**), W.S. (Tanganyika **f**), Oliver Selz (Victoria **f**, *A. burtoni*), Marcel Haesler (*O. niloticus*).

converge in the cichlid visual system, where trichromatic colour vision, eight different opsin genes and novel spherical lenses promote sensitivity in the highly dimensional visual world of clear-water lakes[12–14]. Rapidly evolving sex determination systems, often linked to male and female colour patterns, may also speed cichlid diversification[15,16]. Ecological, social and behavioural variation correlates with striking diversity in brain structures[17] that appears early in development[18].

Exceptional phenotypic variation, even among closely related species, makes cichlids different from most other fish groups, including those that share the same habitats with them but have not diversified as much, as well as those that have radiated into much smaller species flocks in northern temperate lakes[19]. However, how cichlids evolve in this exceptionally highly dimensional phenotype space remains unexplained.

We sequenced the genomes of five representative cichlid species from throughout the East African haplo-tilapiine lineage (Extended Data Fig 1a), which gave rise to all East African cichlid radiations. These five lineages diverged primarily through geographical isolation, and three of them subsequently underwent adaptive radiations in the three largest lakes of Africa (Fig. 1). Here we describe the comparative analyses of the five genomes coupled with an analysis of the genetic basis of species divergence in the Lake Victoria species flock to examine the genomic substrate for rapid evolutionary diversification.

Accelerated gene evolution

To assess whether accelerated sequence evolution was a general feature of East African cichlids, we annotated the genomes of all five cichlids (Extended Data Fig. 1a) and estimated the nonsynonymous/synonymous nucleotide substitution (d*N*/d*S*) ratio by sampling the concatenated alignments of all genes annotated with particular gene ontology (GO) terms. An elevated rate of nonsynonymous nucleotide substitutions can indicate accelerated evolution (either due to relaxed constraint or positive selection); this approach has been applied previously in the context of cichlid vision[13] and morphology[20,21]. We obtained significantly higher d*N*/d*S* ranks in *O. niloticus* (89 terms) compared to stickleback (11 terms), but considerably higher ranks still in the lineages of the East African radiation, haplochromines (299 terms) and *N. brichardi* (254 terms), (Extended Data Fig. 1b). In general, terms involved in morphological and developmental processes ranked significantly higher in haplochromines than in *O. niloticus* (P value = 0.036, Mann–Whitney U-test).

Amongst protein-coding genes with an increased number of nonsynonymous variants in haplochromines compared to *N. brichardi* and *O. niloticus*, two developmental genes, *nog2* and *bmpr1b*, emerged showing haplochromine-specific substitutions. This result is notable given that three genes, a ligand (*bmp4*)[21], a receptor (*bmpr1b*) and an antagonist (*nog2*) in the BMP pathway, all known to influence cichlid jaw morphology, show accelerated rates of protein evolution in haplochromine cichlids.

Of 22 candidate genes previously identified in teleost morphogenesis, vision and pigmentation, three are predicted to have undergone accelerated evolution in the common ancestors of the East African radiations suggesting a role in the diversification of cichlids: endothelin receptor type B1 (*ednrb1*) affects colour patterning[22] and perhaps pharyngeal jaw

development (Extended Data Fig. 2); green-sensitive opsin (*kfh-g*) and Rhodopsin (*rho*) are proteins important in vision.

Gene duplication

Gene duplication allows for subsequent divergent evolution of the resultant gene copies, enabling functional innovation of the proteins and/or expression patterns[23]. East African cichlids, including *Oreochromis niloticus*, possess an unexpectedly large number of gene duplicates. We find 280 duplications in the lineage leading to the common ancestor of the lake radiations and 148 events in the common ancestor of the haplochromines. When normalizing for branch lengths this corresponds to an approximately 4.5- to 6-fold increase in gene duplications that occurred in the common ancestor of the East African lake radiations relative to older clades, and an even higher duplication rate in the common ancestor of just the haplochromines (Fig. 2, Extended Data Fig. 3a–c).

Inferred duplication rates in ancestral populations exceeded those in the extant taxa (Fig. 2). This could reflect the technical challenge of separating young, near-identical gene paralogues or true reduced rates in each lake radiation. Additionally, we could be underestimating lineage-specific rates of duplication owing to the sampling of a single species per radiation, if duplications accumulate during speciation but only some become fixed.

Cichlid-specific gene duplicates do not show statistically significant enrichment for particular gene categories (Supplementary Information). Expansion of the olfactory receptor gene family, which is a frequent feature of vertebrate evolution[24], was also seen in *O. niloticus*, but not in any of the lake cichlids (Extended Data Fig. 4; Supplementary Information). Retained duplicated genes are known to often diverge in function through neo- or subfunctionalization[25], and this has been suggested as part of the reason why bony fish generally are so species-rich (more than 50% of all known species of vertebrates are fish). Moreover, differential retention of alternative copies of duplicated genes through the process of divergent resolution has been suggested to promote speciation rates directly[26].

Differences in the expression patterns of duplicate genes may contribute to evolutionary divergence of species. The expression patterns of 888 duplicate gene pairs from the common ancestor of the East Africa cichlids were categorized according to whether they are expressed widely among tissues (52.8%), are similarly restricted in their expression patterns for both gene copies (26.6%), or, in at least one gene copy, have newly gained expression in one or more tissues (20.6%). 7.5% of duplicates lost or gained complete tissue specificity, many (43%) of which

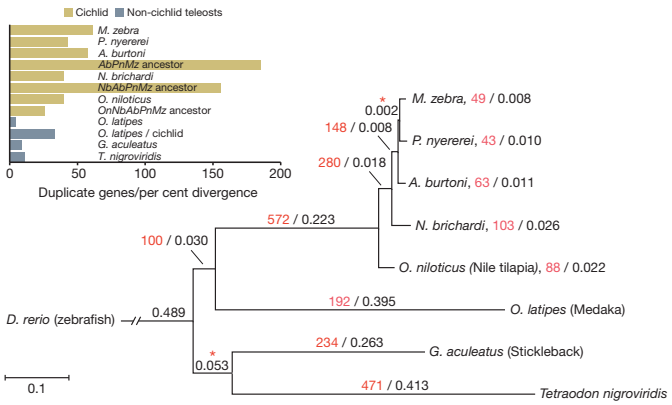


Figure 2 | Gene duplication in the ancestry of East African lake cichlids. Black numbers represents species divergence calculated as neutral genomic divergence between the sequenced species using ~2.7 million fourfold degenerate sites from the alignment of 9 teleost genomes. This neutral substitution model suggests ~2% pairwise divergence between the three haplochromines and a ~6% divergence to *N. brichardi*. Red numbers represent duplicated genes. Asterisks indicate excluded branches owing to incomplete lineage sorting in haplochromines or weak support of consensus species tree.

have gained specific expression in the testis. In each of the stomatin and *RNF141* gene pairs, one gene copy is broadly expressed whereas expression of the other is restricted to the testis (Extended Data Fig. 3d). *RNF141* is the zebrafish orthologue of the human *ZNF230*, a transcription factor suggested to have a role during spermatogenesis. This observation is particularly interesting in the context of strong sexual selection[14] observed in many East African cichlids[15,16], including our sequenced species with the exception of *N. brichardi*.

Transposable element insertions alter gene expression

As in other teleosts, approximately 16–19% of the four East African cichlid genomes consist of transposable elements (TEs), and over 60% of cichlid TEs are DNA transposons (Extended Data Fig. 5; Supplementary Information). Three waves of TE insertions were detected in each of the cichlid genomes (Extended Data Fig. 6a–f), including a cichlid-specific burst of the Tigger family[27]. Notably, this TE family has continued expanding in the youngest radiation, Lake Victoria (Extended Data Fig. 6a).

We analysed the distribution of TE insertions near the 5′ untranslated region (5′ UTR; 0–20 kilobases upstream), or 3′ UTR (0–20 kb downstream) of orthologous gene pairs. We find that genes with TE insertions near the 5′ UTRs are significantly associated with increased gene expression in all tissues (false discovery rate (FDR) < 0.05, Mann–Whitney test, Extended Data Fig. 7a) compared to genes without TE insertions. In contrast, TE insertions near 3′ UTRs are significantly associated with increased gene expression in all tissues except brain and skeletal muscle (FDR < 0.05, Mann–Whitney U-test).

Generally, when inserted within or near genes in the transcriptional sense orientation, TE insertions show the expected pattern of purifying selection. Such TEs often contain polyadenylation signals that result in transcriptional arrest[27]. In all five cichlid species, intronic TE insertions occur preferentially in the antisense orientation of protein-coding genes, with the strongest bias being observed for long terminal repeats (LTRs) or long interspersed nucleotide repetitive elements (LINEs) (Extended Data Fig. 7b). As expected, intronic DNA transposons and LINEs or LTRs present in intergenic regions fail to show a significant orientation bias, and short interspersed nucleotide repetitive elements (SINE) show a moderate bias for sense insertions (Extended Data Fig. 7c).

Surprisingly, none of the five cichlid genomes showed any deficit of sense-oriented LINE insertions with approximately 15% divergence, which correspond to a time of transposable element insertions in the common ancestor of the haplo-tilapiine cichlids (Extended Data Fig. 7d). This suggests that ancestral East African cichlids went through an extended period of relaxed purifying selection during which overall TE activity increased (Extended Data Fig. 6a–f). However, in more recent history, haplochromine cichlids showed an increased efficiency in purging potentially deleterious TE insertions (Extended Data Fig. 7d).

Divergence of regulatory elements

To identify potential regulatory sequences that have diverged among the East African cichlids, we first predicted conserved noncoding elements (CNEs)[28] in Nile tilapia and eight other teleosts using a 9-way alignment of teleost genomes (zebrafish, *Tetraodon*, stickleback, medaka and the five cichlids; Supplementary Information). We then identified 13,053 highly conserved noncoding elements (hCNEs) in tilapia and medaka. These are expected to be similarly conserved among the four East African lake cichlids as they shared a common ancestor with Nile tilapia more recently than with medaka. Among these hCNEs we searched for CNEs that exhibited significant changes (accelerated CNEs, aCNEs) (FDR-adjusted $P < 0.05$). A total of 625 such aCNEs (4.8%) were found to have diverged in one or more of the East African lake cichlids. Whereas the majority of aCNEs (93%) have experienced a higher rate of nucleotide substitutions, approximately a quarter have also experienced insertions (23%) and/or deletions (32%), again suggesting relaxed purifying selection. The aCNEs are distributed in intergenic regions (70%), introns (28%) and UTRs (2%) of protein-coding genes (Supplementary information).

The largest number of aCNEs is found in *N. brichardi* ($n = 214$), with lower numbers found in *A. burtoni* ($n = 140$), *P. nyererei* ($n = 129$) and *M. zebra* ($n = 142$). Approximately 60% of the aCNEs ($n = 370$) are accelerated in only one lineage. The remaining aCNEs have either accumulated mutations independently in several lineages, or their accelerated evolution was initiated in a common ancestor.

The majority of aCNEs in lake cichlids showed enrichment for nearby genes involved in 'homophilic cell adhesion' ($P = 5.8 \times 10^{-4}$) and 'G-protein coupled receptor activity' ($P = 6.4 \times 10^{-4}$). To verify the *cis*-regulatory function of these aCNEs, we assayed the ability of six selected aCNEs and their corresponding *O. niloticus* hCNEs to drive reporter gene expression in transgenic zebrafish. The assays not only indicated their potential to function as enhancers, but also demonstrated that aCNEs have altered the expression pattern compared to their homologous hCNEs, indicating their potential for altering expression of their target genes in a tissue-specific manner. We illustrate this with an example in Extended Data Fig. 8 (additional examples in Extended Data Fig. 9).

Novel microRNAs alter gene expression

MiRNAs offer yet another effective way of altering gene expression programs. We identified 1,344 miRNA loci (259–286 per cichlid species) from deep sequencing of small RNAs in late stage embryos (Extended Data Fig. 10a). By comparing these loci with known teleost microRNAs (Supplementary Information) we discovered: (1) 40 cases of *de novo* miRNA emergence and nine cases of apparent miRNA loss; (2) four distinct mature miRNAs with mutation(s) in the seed sequence; (3) at least 9 cases of *arm* switching[29], (4) one case of seed shifting[29], and (5) 92 distinct miRNAs with mutation(s) outside the seed sequence.

We explored miRNA spatial expression patterns in one case of arm switching (*t_mze-miR-7132a-5p* and *t_mze-miR-7132a-3p)* and for four *de novo* miRNAs (Fig. 3 and Extended Data Fig. 10). In the case of arm switching, spatial expression of the miRNA is clearly differentiated between the two pairs, consistent with results described previously[30]. The spatial expression of the four *de novo* miRNAs (*miR-10029*, *miR10032*, *miR-10044*, *miR-10049*) is confined to specific tissues (for example, fins, facial skeleton, brain) and is strikingly complementary to genes predicted to contain target sites for these miRNAs (*miR-10032* targets *neurod2*, and *miR-10029* targets *bmpr1b*). The *neurod2* gene is known to be involved in brain development and neural differentiation whereas *bmpr1b*, previously described amongst the fast evolving genes, is implicated in the development and morphogenesis of nearly all organ systems.

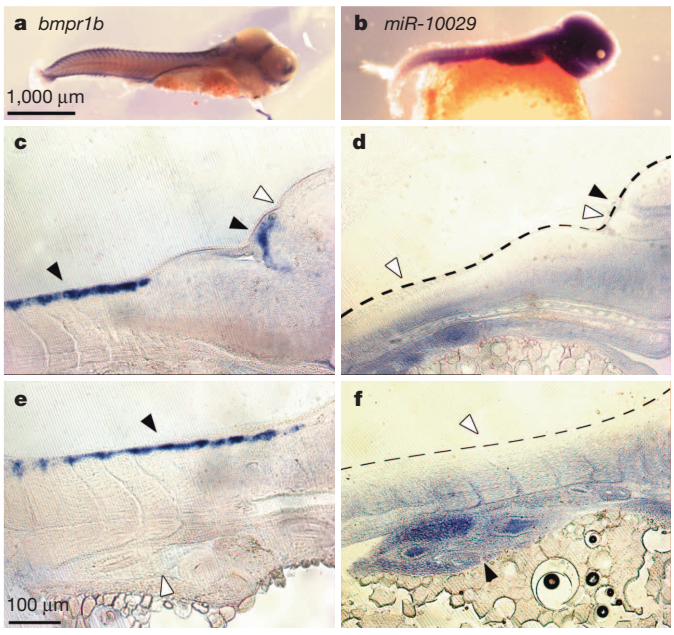


Figure 3 | Novel cichlid microRNAs. **a**–**f**, Complementary expression of *mir-10029* (**b**, **d**, **f**) and its predicted target gene *bmpr1b* (**a**, **c**, **e**) in stage 18 (6 days post-fertilization) *Metriaclima zebra* embryos. **c**–**f** are 18-µm sagittal sections. In **c** and **d** arrows point to expression (black) or lack of expression (white) in the somites, presumptive cerebellum, and optic tectum (from left to right). In **e** and **f**, arrows point to expression and lack of expression in the somites (dorsal) and the gut (ventral). In all panels, anterior is to the right.

Extensive shared polymorphisms

Owing to their relatively recent divergence time and the potential for gene flow between lakes[8,9,31], we predicted widespread incomplete lineage sorting (ILS) among haplochromine cichlids. We found that nearly half (43%) of the nucleotides sequenced are incompletely sorted amongst the three haplochromines (Fig. 4a). Furthermore, assuming a constant mutation rate, and an *A. burtoni–M. zebra–P. nyererei* speciation event ~10 million years ago (Myr ago) (ranging from 7 Myr ago to 15 Myr ago depending on whether Gondwana rifting dates are included or excluded from calibration[32]), we predict the subsequent speciation event between the lineages to which *M. zebra* and *P. nyererei* belong to about 8.5 Myr ago (Supplementary Information). The degree of ILS is highly variable across chromosomes. Compared to intergenic regions, coding regions were found to be slightly, yet significantly, depleted in ILS (43.5% vs 41.0%, $P < 0.001$). Reduction of ILS in coding versus noncoding regions in allopatric lineages of haplochromine cichlids is less than that found in the similarly divergent primate trio, gorilla–chimpanzee–human (30% vs 22%)[33]. This suggests that natural selection has been a more efficient force on primate genomes than on the allopatrically diverging genomes of the haplochromine cichlid lineages, with important implications for genetic diversity in the radiations to which these lineages gave rise.

Lake Victoria, a recent evolutionary radiation

Cichlid fish adaptive radiation is characterized by rapid speciation without geographical isolation. In Lake Victoria, several hundred endemic species emerged within the past 15,000–100,000 years[34]. We analysed patterns of genome-wide genetic variation in six sympatric and closely related species of the genera *Pundamilia*, *Mbipia* and *Neochromis*, all of which are endemic to Lake Victoria. We used the *P. nyererei* genome to investigate the pattern and magnitude of genomic differentiation in pairwise species comparisons. We then further characterized the regions of genomic differentiation to learn about: (1) the genomic distribution of divergent sites putatively under selection; (2) their nature (coding vs regulatory); (3) whether diversification occurred by selection on old standing variation, newer mutations or both.

Divergent selection on many genes

Analyses of restriction-site-associated DNA (RAD) data showed that the average genome-wide divergence was significant in all pairwise species comparisons ($P < 0.001$). In each pairwise comparison, we find many SNPs with high fixation index (F_{ST}) values distributed across all chromosomes (Fig. 4c). In each pair, 250 to 439 of these SNPs constitute significant outliers from the F_{ST} distribution (FDR < 5%; Fig. 4c), and BAYESCAN results indicate numerous loci under selection. Phylogenetic trees reconstructed from the concatenated RAD sequence data resolve species with high bootstrap support[35], and loci putatively under selection play a strong role in differentiating species (Fig. 4b). Taken together, these results suggest that even the most recent rapid speciation in African lake cichlids is associated with genomically widespread divergence. Fixation of alternative alleles between species happens but is restricted to a minority of the many divergent loci, consistent with models of polygenic adaptation from standing genetic variation[36].

We used the annotated *P. nyererei* reference genome to identify genes that diverged during and soon after speciation for three sister species pairs and two pairs of more distant relatives (Fig. 4c). We annotated all SNPs according to their positions in exons and potential *cis*-regulatory elements (in introns and 25 kb either side of genes), and analysed the proportion of SNPs in each category over increasing F_{ST}. In both pairs of sister species that differ primarily in male breeding coloration, the

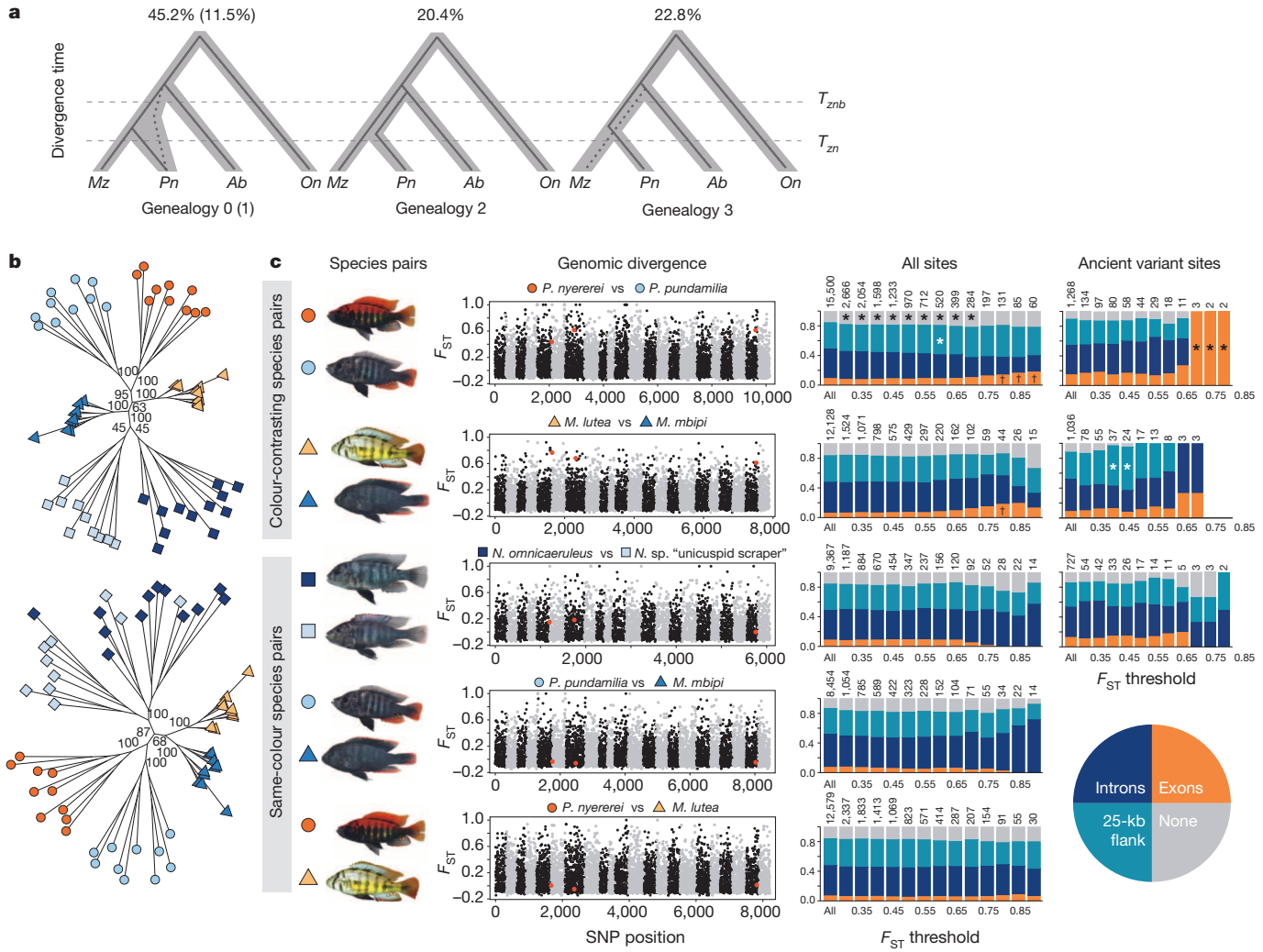


Figure 4 | Genomic divergence stems from incomplete lineage sorting (ILS) and both old and novel coding and noncoding variation. **a**, Coalescence times and trees supporting ILS among the genomes of allopatric East African cichlid lineages were inferred by coalHMM. The most common genealogy matches the known species tree and represents a *M. zebra–P. nyererei* coalescence that falls between the two speciation times, T_{zn} (speciation *M. zebra–P. nyererei*) and T_{znb} (speciation *M. zebra–P. nyererei–A. burtoni*). In genealogies 1 (dashed line), 2, and 3, all coalescence events are ancient and occur before time T_{znb}. **b**, Phylogenetic analysis of RAD-sequence data showing well-supported differentiation among young Victoria species. The complete data set (top) renders the genus *Mbipia* non-monophyletic, exclusion of the top 1% divergent loci (bottom) supports monophyly of each genus. **c**, Genomic divergence in paired comparisons of Lake Victoria cichlids (per-site F_{ST}; black/grey are chromosomes). Sister species from top: *Pundamilia nyererei*/*P. pundamilia* and *Mbipia lutea*/*M. mbipi* differ in male breeding coloration but have conserved morphology; *Neochromis omnicaeruleus*/*N.* sp. "unicuspid scraper" and distant relatives *P. pundamilia*/*M. mbipi* and *P. nyererei*/*M. lutea* have similar coloration but differ in morphology. Red-highlighted SNPs indicate significantly divergent sites between colour-contrasting species, but not between same-colour species. Bar plots show the proportion of SNPs in four annotation categories: exons (orange), introns (dark blue), 25-kb flanking genes (turquoise), or none of the above (grey), for thresholds of increasing F_{ST}. In "All sites" and "Ancient variant sites" analyses, symbols indicate an excess of SNPs in a given annotation category compared to expectations from the full data set or from all non-ancient variant sites, respectively (FDR q-values: $^*q < 0.05$; $\dagger q = 0.05$), (Supplementary Information, Data Portals, Supplementary Population Genomics FTP files).

proportion of SNPs in exons increases from <10% in the full set of SNPs, to >18% at highly divergent SNPs. In the species that have diverged primarily in morphology, we find no exonic variants among highly divergent SNPs, and an increasing proportion of SNPs in introns with increasing F_{ST} (Fig. 4c).

These data suggest contrasting genomic mechanisms underlying phenotypic evolution depending on whether speciation is driven primarily by divergence of coloration and associated traits or by divergence of morphology associated with feeding ecology. This supports two predictions from evolutionary developmental biology[37]: (1) variation in coding sequence is most likely to be involved in the divergence of physiological and/or terminally differentiated traits like colour; (2) regulatory variation is more important in morphological changes involving genes that have pleiotropic effects in developmental networks.

For the *Pundamilia* species pair, putative regulatory SNPs with F_{ST} values significantly greater than zero show enrichment in conserved transcription factor binding sites and PhastCon elements (conserved elements across 46 vertebrate species), supporting a regulatory role for these variants. GO term enrichment analyses indicate that exonic SNPs are associated with metabolism and biosynthesis processes, while putative regulatory SNPs are associated with terms related to morphogenesis and development.

Comparing F_{ST} for each SNP in all six pairwise comparisons of the *Mbipia* and *Pundamilia* species revealed 3 candidate regulatory SNPs on LG6, 7 and 22 that are highly divergent in all comparisons of species with different colours, but not significantly differentiated between species with similar colours (Fig. 4c). The SNP on LG7 falls within a known quantitative trait locus (QTL) interval for yellow versus blue colour (and

sex determination) in Malawi cichlids[15]. None of these SNPs are fixed differences between species, suggesting polygenic adaptation.

Sorting of ancient polymorphisms

To investigate whether ancient genetic variation, predating the origin of the Lake Victoria species flock, was an important source of alleles that are divergently sorted during speciation, for SNPs in each of the three Victoria sister species pair comparisons, we identified orthologous sites among the four other cichlid genomes. We find 14–15% of all Victoria SNPs are also variable among the other cichlid genomes. Among these 'ancient variants', the proportion of SNPs in exons increases from 9–15% among all sites to 30–100% at highly divergent SNPs in both pairs of sister species that differ primarily in male breeding coloration (Fig. 4c). Among the ancient exonic variants that became fixed in the red/blue *Pundamilia* speciation event is *srd5a2b*, a teleost-specific duplicate of *srd5a2* which, in mammals, converts testosterone to dihydrotestosterone and has been implicated in sexual differentiation[38]. In the blue sister species that have diverged primarily in morphology, two ancient variants in potential *cis*-regulatory regions are highly divergent despite incomplete reproductive isolation among these incipient species[39] (Fig. 4b). We compared the proportions of putative ancient variants to all SNPs between annotation categories, and find evidence for higher proportions of ancient variants in gene-associated regions than in non-genic regions (likelihood ratio tests on 2×2 contingency tables; exons: *Pundamilia* $P = 0.016$, *Neochromis* $P = 0.015$; flanking regions: *Pundamilia* $P = 0.020$; all other $P > 0.1$).

These analyses suggest that the genomic substrate for adaptive radiation includes ample coding and regulatory polymorphism, likely to be present well before the start of the radiations, some of which became subsequently sorted during species divergence.

Conclusions

In African lakes, nearly 1,500 new species of cichlid fish evolved in a few million years when environmentally determined opportunity for sexual selection and ecological niche expansion[4] was met by an evolutionary lineage with unusual potential to adapt, speciate and diversify. Our analyses of five cichlid species representing five different lineages in the haplo-tilapiine clade, some of which gave rise to radiations, and of six closely related species from the most recent radiation, shed light into the complex genomic mechanisms that may give East African cichlids their unusual propensity for diversification.

We provide evidence for accumulation of genetic variation under relaxed constraint preceding radiation and involving multiple evolutionary mechanisms, including accelerated evolution of regulatory and coding sequence, increased gene duplication, TE insertions, novel micoRNAs and retention of ancient polymorphisms, possibly including interspecific hybridization. In addition, our data on genomic divergence within the Lake Victoria species flock suggest that adaptive radiation within the lakes is associated with divergent selection on many regions in the genome, both coding and regulatory, often recruiting old alleles from standing variation.

We conclude that neutral and adaptive processes both make important contributions to the genetic basis of cichlid radiations, but their roles are distinct and their relative importance has changed through time: neutral (and non-adaptive) processes seem to have been crucial to amassing genomic variation, whereas selection subsequently sorted some of this variation. The interaction of both is likely to have been necessary for generating many and diverse new species in very short periods of time.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 18 December 2013; accepted 1 August 2014.
Published online 3 September 2014.

1. Darwin, C. *On the Origin of Species* 6th edn (John Murray, 1859).
2. Simpson, G. G. *Tempo and Mode in Evolution* (Columbia Univ. Press, 1944).
3. Kocher, T. D. Adaptive evolution and explosive speciation: the cichlid fish model. *Nature Rev. Genet.* **5,** 288–298 (2004).
4. Wagner, C. E., Harmon, L. J. & Seehausen, O. Ecological opportunity and sexual selection together predict adaptive radiation. *Nature* **487,** 366–369 (2012).
5. Meyer, A. Morphometrics and allometry in the trophically polymorphic cichlid fish, *Cichlasoma citrinellum*: alternative adaptations and ontogenic changes in shape. *J. Zool.* **221,** 237–260 (1990).
6. Cohen, A. S., Soreghan, M. J. & Schloz, C. A. Estimating the age of formation of lakes: an example from Lake Tanganyika, East African Rift system. *Geology* **21,** 511–514 (1993).
7. McCune, A. *How Fast is Speciation: Molecular, Geological and Phylogenetic Evidences from Adaptive Radiations of Fish* pp. 585–610 (Cambridge Univ. Press, 1997).
8. Joyce, D. A. *et al.* Repeated colonization and hybridization in Lake Malawi cichlids. *Curr. Biol.* **21,** R108–R109 (2011).
9. Loh, Y.-H. E. *et al.* Origins of shared genetic variation in african cichlids. *Mol. Biol. Evol.* **30,** 906–917 (2013).
10. Albertson, R. C., Streelman, J. T., Kocher, T. D. & Yelick, P. C. Integration and evolution of the cichlid mandible: the molecular basis of alternate feeding strategies. *Proc. Natl Acad. Sci. USA* **102,** 16287–16292 (2005).
11. Muschick, M., Barluenga, M., Salzburger, W. & Meyer, A. Adaptive phenotypic plasticity in the Midas cichlid fish pharyngeal jaw and its relevance in adaptive radiation. *BMC Evol. Biol.* **11,** 116 (2011).
12. Fernald, R. D. Vision and behavior in an african cichlid fish. *Am. Sci.* **72,** 58–65 (1984).
13. Hofmann, C. M. *et al.* The eyes have it: regulatory and structural changes both underlie cichlid visual pigment diversity. *PLoS Biol.* **7,** e1000266 (2009).
14. Maan, M. E. *et al.* Intraspecific sexual selection on a speciation trait, male coloration, in the Lake Victoria cichlid *Pundamilia nyererei*. *Proc. R. Soc. Lond. B* **271,** 2445–2452 (2004).
15. Parnell, N. F. & Streelman, J. T. Genetic interactions controlling sex and color establish the potential for sexual conflict in Lake Malawi cichlid fishes. *Heredity* **110,** 239–246 (2013).
16. Roberts, R. B., Ser, J. R. & Kocher, T. D. Sexual conflict resolved by invasion of a novel sex determiner in Lake Malawi cichlid fishes. *Science* **326,** 998–1001 (2009).
17. Huber, R., vanStaaden, M. J., Kaufman, L. S. & Liem, K. F. Microhabitat use, trophic patterns, and the evolution of brain structure in African cichlids. *Brain Behav. Evol.* **50,** 167–182 (1997).
18. Sylvester, J. B. *et al.* Competing signals drive telencephalon diversity. *Nat. Commun.* **4,** 1745 (2013).
19. Jones, F. C. *et al.* The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484,** 55–61 (2012).
20. Fan, S., Elmer, K. R. & Meyer, A. Positive Darwinian selection drives the evolution of the morphology-related gene, EPCAM, in particularly species-rich lineages of African cichlid fishes. *J. Mol. Evol.* **73,** 1–9 (2011).
21. Terai, Y., Morikawa, N. & Okada, N. The evolution of the pro-domain of bone morphogenetic protein 4 (Bmp4) in an explosively speciated lineage of East African cichlid fishes. *Mol. Biol. Evol.* **19,** 1628–1632 (2002).
22. Parichy, D. M. *et al.* Mutational analysis of *endothelin receptor b1* (*rose*) during neural crest and pigment pattern development in the zebrafish *Danio rerio*. *Dev. Biol.* **227,** 294–306 (2000).
23. Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* **290,** 1151–1155 (2000).
24. Plenderleith, M., van Oosterhout, C., Robinson, R. L. & Turner, G. F. Female preference for conspecific males based on olfactory cues in a Lake Malawi cichlid fish. *Biol. Lett.* **1,** 411–414 (2005).
25. Taylor, J. S., Braasch, I., Frickey, T., Meyer, A. & Van de Peer, Y. Genome duplication, a trait shared by 22000 species of ray-finned fish. *Genome Res.* **13,** 382–390 (2003).
26. Taylor, J. S., Van de Peer, Y. & Meyer, A. Genome duplication, divergent resolution and speciation. *Trends Genet.* **17,** 299–301 (2001).
27. Medstrand, P., van de Lagemaat, L. N. & Mager, D. L. Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome Res.* **12,** 1483–1495 (2002).
28. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050 (2005).
29. Berezikov, E. Evolution of microRNA diversity and regulation in animals. *Nature Rev. Genet.* **12,** 846–860 (2011).
30. Ro, S., Park, C., Young, D., Sanders, K. M. & Yan, W. Tissue-dependent paired expression of miRNAs. *Nucleic Acids Res.* **35,** 5944–5953 (2007).
31. Salzburger, W., Meyer, A., Baric, S., Verheyen, E. & Sturmbauer, C. Phylogeny of the Lake Tanganyika cichlid species flock and its relationship to the Central and East African haplochromine cichlid fish faunas. *Syst. Biol.* **51,** 113–135 (2002).
32. Genner, M. J. *et al.* Age of cichlids: New dates for ancient lake fish radiations. *Mol. Biol. Evol.* **24,** 1269–1282 (2007).
33. Scally, A. *et al.* Insights into hominid evolution from the gorilla genome sequence. *Nature* **483,** 169–175 (2012).
34. Johnson, T. C. *et al.* Late Pleistocene desiccation of Lake Victoria and rapid evolution of cichlid fishes. *Science* **273,** 1091–1093 (1996).
35. Wagner, C. E. *et al.* Genome-wide RAD sequence data provides unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Mol. Ecol.* **22,** 787–798 (2012).
36. Barrett, R. D. & Schluter, D. Adaptation from standing genetic variation. *Trends Ecol. Evol.* **23,** 38–44 (2008).
37. Stern, D. L. & Orgogozo, V. The loci of evolution: how predictable is genetic evolution? *Evolution* **62,** 2155–2177 (2008).

38. Thigpen, A. E. *et al.* Molecular genetics of steroid 5 alpha-reductase 2 deficiency. *J. Clin. Invest.* **90,** 799–809 (1992).
39. Magalhaes, I. S., Lundsgaard-Hansen, B., Mwaiko, S. & Seehausen, O. Evolutionary divergence in replicate pairs of ecotypes of Lake Victoria cichlid fish. *Evol. Ecol. Res.* **14,** 381–401 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We would like to thank the Broad Institute Genomics Platform for sequencing of the 5 cichlid genomes and transcriptomes. Sequencing, assembly, annotation and analysis by Broad Institute were supported by grants from the National Human Genome Research Institute (NHGRI). Genome evolution, duplication and TE analysis, ILS and ancient variant analyses were also supported by Swiss National Science Foundation grant PBLAP3-142774 awarded to D.B. and by University of Oxford Nuffield Department of Medicine Prize Studentship to Y.I.L. TE and copy number variation analyses were supported by the German Science Foundation (DFG), and advanced grant 29700 (''GenAdap'') by the European Research Council (ERC). CNE analysis and zebrafish functional assays were supported by the Biomedical Research Council of A*STAR, Singapore. MicroRNA sequencing and annotation was supported by ERC Starting Grant to E.A.M.; M.M. was supported by a fellowship from the Wellcome Trust. MicroRNA and target *in situ* hybridization was supported by grant 2R01DE019637-04 to J.T.S. Population genomics analyses were supported by Swiss National Science Foundation grants 31003A-118293 and 31003A-144046 to O.S.

Author Contributions T.D.K., R.D.F., A.M., O.S., J.T.S., K.L.C., N.O., J.-F.B., D.J.P. and H.A.H. conceived the original tilapia white paper. F.D.P. , K.L.-T. and E.S.L. revised, planned and oversaw the genome project. D.J.P., W.S., H. S. G., M.E.S., O.S., K.L.C., T.D.K., G.H., O.E. and H.A.H. provided tissues and RNAs for sequencing. C.A. prepared the high molecular weight tilapia DNA. M.L. extracted genomic DNA for sequencing. L.W. prepared 40-kb libraries (Fossils) for Illumina sequencing. R.S. performed quality control of RNA. J.A., J.J. and F.D.P. oversaw the sequencing and assembly of genomes and transcriptomes as well as submissions of data. J.T.M. and P.R. performed quality control of assemblies and alignments of genomes. J.M.T. performed *de novo* assembly of transcriptomes. M.C. performed quality control of tilapia and *M. zebra* assemblies. A.B., Sa.Y., I.M., S.G., D.P., F.J.R., T.S., Sh.Y. and D.B.J. assembled the genome. F.G., R.G., M.R., J.-F.B., H.D'C., C.O.-C. contributed to the tilapia radiation hybrid map. F.B.-H. and N.A. analysed the *OR* and *TAAR* gene families. B.A., T.H. and S.S. annotated the tilapia genome. D.B. and Y.I.L. annotated the *N. brichardi* and the lake cichlids. D.B. performed gene expression, genome evolution, gene duplication and TE insertion analyses. Y.I.L. and L. S.-P. performed quality control of RNA-seq data and assemblies, gene evolution, incomplete lineage sorting and ancient variant analyses. S.F., Oleg S. and A.M., N.O., M.N. and H.N. analysed the TE landscape of cichlid genomes. S.F., Oleg S. and A.M. performed the TE burst history analysis and analysed copy number variants using read depth. E.B. and S.C.P.R. analysed duplications by comparative genomic hybridization (aCGH). H.A.H. and R.M.H. performed PCR to validate the transcriptome. A.Y.N., Z.W.L., A.P.L. and B.V. performed conserved CNE analysis and functional assays of cichlid CNEs. M.M. and E.M. performed microRNA sequencing and annotation from embryos of cichlid species as well as target identification. R.A., F.J.T. and R.D.F. annotated adult brain microRNAs in *A. burtoni*. R.B., N.S.H. and J.T.S. performed microRNA and target gene *in situ* hybridization. O.S. designed and oversaw the population genomics data analysis from Lake Victoria species; L.G., S.M. and I.K. generated the data; C.E.W., I.K., H.J.N. and O.S. analysed the data. F.D.P., K.L.-T. and O.S. wrote the manuscript with input from D.B., C.E.W. and Y.I.L., I.K., J.T.S., W.H., C.P.P. as well as additional authors. L.G. assisted with figure preparation and coordination.

Author Information Genome assemblies and transcriptomes have been deposited in GenBank. The BioProject Identifiers are as follows. Genome sequencing: PRJNA59571 (SRP004171) for *O. niloticus*; PRJNA60365 (SRP004799) for *N. brichardi*; PRJNA60367 (SRP004869) for *P. nyererei*; PRJNA60369 (SRP004788) for *M. zebra*; and PRJNA60363 (SRP004787) for *A. burtoni.* Transcriptome sequencing (mRNAs): PRJNA78915 for *O. niloticus*; PRJNA77747 for *N. brichardi*; PRJNA83153 for *P. nyererei*; PRJNA77743 for *M. zebra*; and PRJNA78185 for *A. burtoni.* Additional SRA information for each tissue can be found in the Supplementary Informations. Transcriptome sequencing (microRNAs): PRJNA221867 (SRS489376) for *O. niloticus*; PRJNA222491 (SRS491903) for *N. brichardi*; PRJNA222489 (SRS491906) for *P. nyererei*; PRJNA221871 (SRS491904) for *M. zebra*; and PRJNA222490 (SRS491905) for *A. burtoni.* Cichlid microRNAs were deposited in miRBase. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to F.D.P. (Federica.di-palma@tgac.ac.uk), K.L.-T. (Kersli@broadinstitute.org), J.T.S. (todd.streelman@biology.gatech.edu), and O.S. (ole.seehausen@eawag.ch).

ARTICLE

doi:10.1038/nature13438

Proteogenomic characterization of human colon and rectal cancer

Bing Zhang[1,2], Jing Wang[1], Xiaojing Wang[1], Jing Zhu[1], Qi Liu[1], Zhiao Shi[3,4], Matthew C. Chambers[1], Lisa J. Zimmerman[5,6], Kent F. Shaddox[6], Sangtae Kim[7], Sherri R. Davies[8], Sean Wang[9], Pei Wang[10], Christopher R. Kinsinger[11], Robert C. Rivers[11], Henry Rodriguez[11], R. Reid Townsend[8], Matthew J. C. Ellis[8], Steven A. Carr[12], David L. Tabb[1], Robert J. Coffey[13], Robbert J. C. Slebos[2,6], Daniel C. Liebler[5,6] & the NCI CPTAC*

Extensive genomic characterization of human cancers presents the problem of inference from genomic abnormalities to cancer phenotypes. To address this problem, we analysed proteomes of colon and rectal tumours characterized previously by The Cancer Genome Atlas (TCGA) and perform integrated proteogenomic analyses. Somatic variants displayed reduced protein abundance compared to germline variants. Messenger RNA transcript abundance did not reliably predict protein abundance differences between tumours. Proteomics identified five proteomic subtypes in the TCGA cohort, two of which overlapped with the TCGA 'microsatellite instability/CpG island methylation phenotype' transcriptomic subtype, but had distinct mutation, methylation and protein expression patterns associated with different clinical outcomes. Although copy number alterations showed strong *cis*- and *trans*-effects on mRNA abundance, relatively few of these extend to the protein level. Thus, proteomics data enabled prioritization of candidate driver genes. The chromosome 20q amplicon was associated with the largest global changes at both mRNA and protein levels; proteomics data highlighted potential 20q candidates, including *HNF4A* (hepatocyte nuclear factor 4, alpha), *TOMM34* (translocase of outer mitochondrial membrane 34) and *SRC* (SRC proto-oncogene, non-receptor tyrosine kinase). Integrated proteogenomic analysis provides functional context to interpret genomic abnormalities and affords a new paradigm for understanding cancer biology.

TCGA has characterized the genomic features of human cancers[1–6] and this has presented a new challenge of explaining how genomic alterations drive cancers[7]. As proteins link genotypes to phenotypes, the Clinical Proteomic Tumour Analysis Consortium (CPTAC) is performing proteomic analyses of TCGA tumour specimens for selected cancer types. Here we present the first integrated proteogenomic characterization of human cancer with an analysis of the TCGA colorectal cancer (CRC) specimens[6].

The TCGA study affirmed well-established genomic features of CRC and described three transcriptional subtypes, 17 chromosomal regions of significant focal amplification and 28 regions of significant focal deletion, and linked genomic features of CRC to critical signalling pathways. The drivers underlying copy number alterations (CNAs) and transcriptional subtypes are largely unknown, and an integrative analysis of both genomic and proteomic data may provide a more comprehensive understanding of the information flow from DNA to protein to phenotype.

Peptide and protein identification

We performed liquid chromatography–tandem mass spectrometry (LC-MS/MS)-based shotgun proteomic analyses on 95 TCGA tumour samples (Extended Data Fig. 1 and Methods), the clinical and pathological characteristics and TCGA data sets for which are summarized in Supplementary Table 1. Benchmark quality control samples from one basal and one luminal human breast tumour xenograft were analysed in alternating order after every five CRC samples (Methods).

We identified a total of 124,823 distinct peptides among the 95 samples, corresponding to 6,299,756 spectra in an assembly of 7,526 protein groups with a protein-level false discovery rate (FDR) of 2.64% (Methods and Extended Data Fig. 2). To facilitate integration between genomic and proteomic data, a gene-level assembly of the peptides identified 7,211 genes.

A fundamental question in proteogenomics is which protein coding alterations are expressed at the protein level. As standard database search approaches cannot identify variant peptides from MS/MS data, we also performed database searches with customized sequence databases from matched RNA sequencing (RNA-seq) data for individual samples[8,9] (Methods and Extended Data Fig. 3).

We identified 796 single amino acid variants (SAAVs) across all 86 tumours for which matched RNA-seq data were available (Fig. 1a, b and Supplementary Tables 2 and 3), among which 64 corresponded to somatic variants reported by TCGA and 101 were reported in the COSMIC database (that is, COSMIC-supported variants). Of the remaining SAAVs, 526 were listed in the Single Nucleotide Polymorphism database (dbSNP) (that is, dbSNP-supported variants) and are likely to be germline variants. The 162 previously unreported SAAVs might be explained by novel somatic or germline variants, RNA editing, or, in some cases, false discovery.

The identified somatic variants were clearly enriched in the hypermutated samples, whereas the germline variants showed no association with hypermutation (Fig. 1a). Although 58% of the germline variants occurred in two or more samples, almost all somatic variants occurred

[1]Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, Tennessee 37232, USA. [2]Department of Cancer Biology, Vanderbilt University School of Medicine, Nashville, Tennessee 37232, USA. [3]Advanced Computing Center for Research and Education, Vanderbilt University, Nashville, Tennessee 37232, USA. [4]Department of Electrical Engineering and Computer Science, Vanderbilt University, Tennessee 37232, USA. [5]Department of Biochemistry, Vanderbilt University School of Medicine, Nashville, Tennessee 37232, USA. [6]Jim Ayers Institute for Precancer Detection and Diagnosis, Vanderbilt-Ingram Cancer Center, Nashville, Tennessee 37232, USA. [7]Directorate of Fundamental and Computational Sciences, Pacific Northwest National Laboratory, Richland, Washington 99352, USA. [8]Department of Internal Medicine, Washington University School of Medicine, St. Louis, Missouri 63110, USA. [9]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, M2-B500, Seattle, Washington 98109, USA. [10]Department of Genetics and Genomic Sciences, Icahn Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1498, New York, New York 10029, USA. [11]Office of Cancer Clinical Proteomics Research, National Cancer Institute, Bethesda, Maryland 20892, USA. [12]Broad Institute of MIT and Harvard, Cambridge, Maryland 02142, USA. [13]Department of Medicine, Vanderbilt University School of Medicine, Nashville, Tennessee 37232, USA.
*Lists of participants and their affiliations appear at the end of the paper.

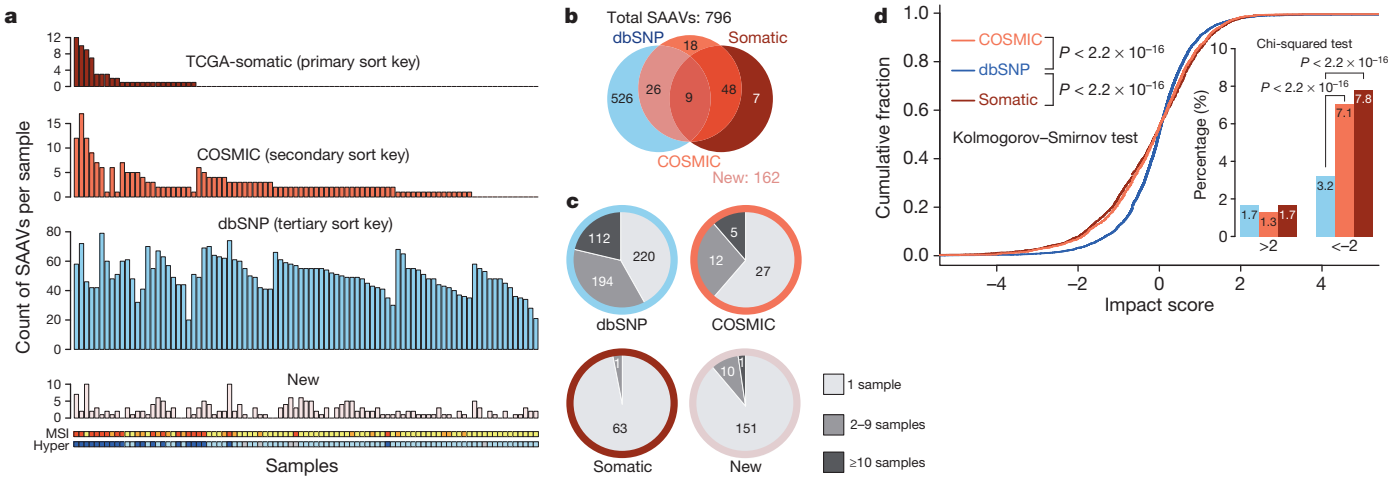


Figure 1 | Summary of detected single amino acid variants (SAAVs) and the impact of single nucleotide variants (SNVs) on protein abundance. **a**, The number of different types of SAAVs (TCGA-reported somatic variants, COSMIC-supported variants, dbSNP-supported variants and new variants) in individual tumour samples. The samples are ordered by the number of detected somatic variants, then COSMIC-supported variants, and then dbSNP-supported variants. The MSI and hypermutation (Hyper) status are labelled below the bar charts for each sample (red, MSI-high; orange, MSI-low; yellow, microsatellite stable; blue, hypermutated; light blue, non-hypermutated; grey, no data). The number of somatic variants and COSMIC-supported variants were significantly higher in MSI-high and hypermutated tumours, whereas the other two types of SAAVs were randomly distributed across the data set. **b**, The total numbers for different types of SAAVs and their overlapping relations. All 796 detected SAAVs were annotated based on previous reports in dbSNP (left circle), COSMIC (middle circle) or TCGA-reported somatic variants (right circle), and their overlapping relations are shown in the Venn diagram. There are 162 SAAVs that have not been reported previously in these databases (new). **c**, Distribution of the frequency of occurrence for different types of SAAVs. Border colours of the pie charts correspond to different SAAV types using the same colour scheme as in **a**. Whereas 58% of dbSNP-supported variants occurred in two or more samples, almost all somatic variants occurred in only one sample each. **d**, SNVs detected in RNA-seq data were separated into three categories (dbSNP-supported, COSMIC-supported and TCGA-somatic). The impact of individual SNVs on protein abundance was calculated (see Methods) and the impact scores for different categories of SNVs were plotted as cumulative fraction curves with two-sided *P* values from the Kolmogorov–Smirnov test labelled. The percentage of SNVs with an absolute impact score greater than 2 was also plotted as an inset, with *P* values from the Chi-squared test. Sample size for the dbSNP-supported, COSMIC-supported and TCGA-somatic variants were 1,2184, 7,492 and 3,302, respectively.

in only one sample (Fig. 1c). The low identification rate for somatic variants may reflect relatively low sequence coverage in shotgun proteomics; however, somatic variants also might negatively impact protein abundance, possibly by reducing translational efficiency or protein stability[10]. Using the protein abundance quantification method described below and detailed in the Methods, we found that somatic variants exerted a significantly stronger negative impact on protein abundance than did dbSNP-supported variants ($P < 2.2 \times 10^{-16}$, Kolmogorov–Smirnov test; Fig. 1d and Methods). The percentage of variants with an impact score of less than −2 was doubled for somatic variants compared to dbSNP-supported variants ($P < 2.2 \times 10^{-16}$, Chi-squared test; Fig. 1d).

Cancer-related variant proteins may serve as candidate protein biomarkers or therapeutic targets. The 108 somatic or COSMIC-supported protein variants mapped to 105 genes, including known cancer genes in the Cancer Gene Census database such as *KRAS*, *CTNNB1*, *SF3B1*, *ALDH2* and *FH*. The list also included 14 targets of FDA-approved drugs or drugs in clinical trials[4], such as *ALDH2*, *HSD17B4*, *PARP1*, *P4HB*, *TST*, *GAK*, *SLC25A24* and *SUPT16H*. A subset of variant peptide sequences, including *KRAS*(Gly12Asp) were verified by targeted analyses of tumour lysates spiked with synthetic, isotope-labelled peptide standards (Methods). One example is shown in Extended Data Fig. 4.

Quantification of protein abundance

To quantify protein abundance, we used spectral counts, which are the total number of MS/MS spectra acquired for peptides from a given protein[11] (Methods and Supplementary Table 4). Analysis of data from benchmark quality control samples demonstrated platform reproducibility throughout the analyses and enabled evaluation of data normalization methods (Extended Data Fig. 5a, b). Based on the minimal spectral count requirement established using the quality control data set (Extended Data Fig. 5c), 3,899 genes with a protein-level FDR of 0.43% were used to compare relative protein abundance across tumour samples.

mRNA versus protein abundance

The matched proteomic and RNA-seq measurements from the TCGA CRC tumours enabled the first global analysis of transcript–protein relationships in a large human tumour cohort (Methods). First, we compared the steady state mRNA and protein abundance for each gene within individual samples (Methods and Extended Data Fig. 6a). All samples showed significant positive mRNA–protein correlation (multiple-test adjusted $P < 0.01$, Spearman's correlation coefficient) and the average correlation between steady state mRNA and protein abundance in individual samples was 0.47 (Fig. 2a), which is comparable to previous reports in multi-cellular organisms[12].

Next, we examined the concordance between mRNA and protein variation of individual genes across the 87 tumours for which 3,764 genes had both mRNA and protein measurements suitable for relative abundance comparison (Methods). Although 89% of the genes showed a positive mRNA–protein correlation, only 32% had statistically significant correlations (Fig. 2b). The average Spearman's correlation between mRNA and protein variation was 0.23, which was comparable to reported values for yeast, mouse and human cell lines[13–15].

To test whether the concordance between protein and mRNA variation is related to the biological function of the gene product, we performed KEGG (Kyoto Encyclopedia of Genes and Genomes) enrichment analysis (Methods and Supplementary Table 5). Genes involved in several metabolic processes showed concordant mRNA and protein variation, whereas other gene classes showed low or even negative concordance in mRNA and protein variation (Fig. 2c). We also found that genes with stable mRNA and stable protein tend to have higher mRNA–protein correlation than those with unstable mRNA and unstable protein ($P = 5.27 \times 10^{-6}$, two-sided Wilcoxon rank-sum test, Methods, Extended Data Fig. 6b). Thus, mRNA measurements are poor predictors of protein abundance variations and both biological functions of the gene products and mRNA and protein stability may govern mRNA–protein correlation.

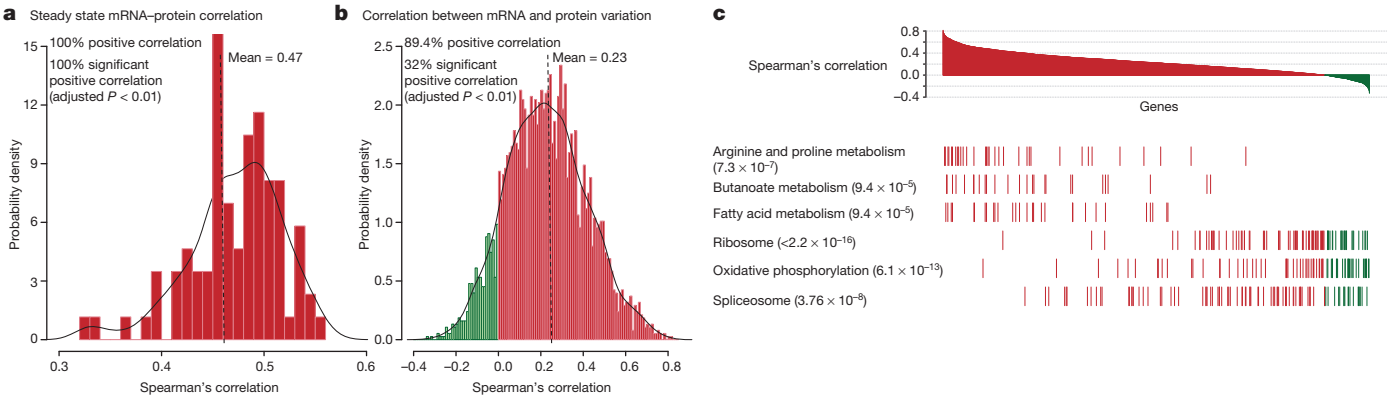


Figure 2 | Correlations between mRNA and protein abundance in TCGA tumours. **a**, Steady state mRNA and protein abundance were positively correlated in all 86 samples (multiple-test adjusted $P < 0.01$) with a mean Spearman's correlation coefficient of 0.47. **b**, mRNA and protein variation were positively correlated for most (89.4%) mRNA-protein pairs across the 87 samples, but only 32% showed significant correlation (multiple-test adjusted $P < 0.01$), with a mean Spearman's correlation coefficient of 0.23. **c**, mRNA and protein levels displayed dramatically different correlation for genes involved in different biological processes. Genes encoding intermediary metabolism functions showed high mRNA–protein correlations, whereas genes involved in oxidative phosphorylation, RNA splicing and ribosome components showed low or negative correlations. Multiple-test adjusted two-sided *P* values from the Kolmogorov–Smirnov test were provided in the parentheses following the KEGG pathway names. Red and green in the figures indicate positive and negative correlations, respectively.

Impact of copy number alterations

The study by TCGA identified 17 regions of significant focal amplification and 28 regions of significant focal deletion. We examined the impact of CNAs on mRNA and protein abundance, including both *cis*-effects on the abundance of genes in the same loci and *trans*-effects on the abundance of genes at other loci in the genome (Methods).

For all 23,125 genes with a CNA measurement in the TCGA data set, we calculated Spearman's correlation with mRNA and protein abundance, respectively for the 3,764 genes with both mRNA and protein measurements (Methods). Examination of the matrix visualizing significant CNA–mRNA correlations (multiple-test adjusted $P < 0.01$) revealed strong positive correlations along the diagonal (Fig. 3a), suggesting strong *cis*-effects of CNAs on mRNA abundance. Most of the diagonal signals corresponded to previously reported arm-level changes[6]. In contrast, the diagonal pattern was much weaker for CNA–protein correlations (Fig. 3b).

To investigate further the *cis*-effects of CNAs, we separated all genes with CNA, mRNA and protein measurements into those in focal amplification regions, focal deletion regions and non-focal regions (that is, chromosomal regions without focal amplification or deletion). As shown in Extended Data Fig. 7, CNA–mRNA correlations were significantly higher than CNA–protein correlations for genes in all three groups ($P < 1.0 \times 10^{-10}$, Kolmogorov–Smirnov test). Moreover, genes in the focal amplification regions showed significantly higher CNA–mRNA and CNA–protein correlations than genes in the non-focal regions ($P = 4.4 \times 10^{-4}$ and 0.02, respectively, Kolmogorov–Smirnov test). However, the same trend was not observed for genes in the focal deletion regions. Therefore, focal amplifications have the strongest *cis*-effects on both mRNA and protein abundance, suggesting that selection for high protein abundance may drive CNA in regions of focal amplification. However, many CNA-driven mRNA level increases do not translate into increased abundance of the corresponding proteins.

Figure 3a, b also revealed multiple *trans*-acting CNA hot spots, defined as chromosomal loci whose alteration is significantly associated with abundance changes of many transcripts or proteins at other loci. Chromosomes 20q, 18, 16, 13 and 7 contained the five strongest hot spots driving global mRNA abundance variation. These hot spots also were strongest at the protein level. Most hot-spot-related transcript changes did not propagate to the protein level, presumably reflecting buffering of protein abundance by post-transcriptional regulation[16,17]. Notably, many hot-spot-associated protein-level alterations occurred in the absence of corresponding mRNA alterations, suggesting that the same *trans*-acting hot spot may exert independent effects at both the transcriptome and proteome levels.

The 20q amplification was associated with the largest global changes in both mRNA and protein levels in this univariate analysis. The same conclusion was reached with a regularized multivariate regression analysis method, remMap[18] (Methods and Supplementary Tables 6–9). These

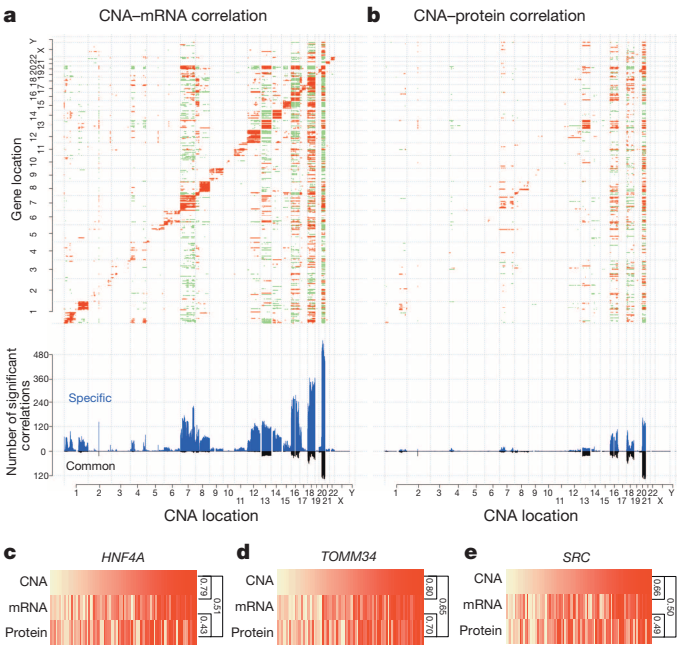


Figure 3 | Effects of copy number alterations on mRNA and protein abundance. **a**, **b**, The top panels show copy-number-abundance correlation matrices for mRNA abundance (**a**) and protein abundance (**b**) with significant positive and negative correlations (multiple-test adjusted $P < 0.01$, Spearman's correlation coefficient) indicated by red and green, respectively, and genes ordered by chromosomal location on both *x* and *y* axes. The bottom panels show the frequency of mRNAs and proteins associated with a particular copy number alteration, where blue and black bars represent associations specific to mRNA and protein or common to both mRNA and protein, respectively. **c**–**e**, *HNF4A* (**c**), *TOMM34* (**d**) and *SRC* (**e**) showed significant CNA–mRNA, mRNA–protein, and CNA–protein correlations (Spearman's correlation coefficient). The colour grade from light yellow to red indicates relatively low-level (yellow) to high-level (red) of copy number, mRNA abundance, or relative protein abundance among the 85 samples, which were ordered by copy number data.

data highlight the importance of 20q amplification in CRC, which has not been well documented in previous studies. Among the 79 genes in the 20q region with quantifiable protein measurements, 67 (85%) showed significant CNA–mRNA correlation, but only 40 (51%) showed significant CNA–protein correlation (multiple-test adjusted $P < 0.01$, Spearman's correlation coefficient, Supplementary Table 10).

As significant CNA–protein correlations identify amplified sequences that translate to high protein abundance, proteomic measurements can help prioritize genes in amplified regions for further examination. Of particular interest among the 40 genes is *HNF4A* (Fig. 3c), a candidate driver gene nominated by TCGA for the 20q13.12 focal amplification peak[6]. HNF4α is a transcription factor with a key role in normal gastrointestinal development[19] and is increasingly being linked to CRC[20]. However, there are contradictory reports on whether HNF4α acts as an oncogene or a tumour suppressor gene in CRC[20]. Upon reanalysis of the *HNF4A* short hairpin RNA (shRNA) knockdown data for CRC cell lines from the Achilles project[21], we found that the dependency of CRC cells on HNF4α correlated significantly with the amplification level of *HNF4A* (Methods and Extended Data Fig. 8), which may partially explain the contradictory roles reported for HNF4α in CRC. Other interesting candidates included *TOMM34* (Fig. 3d), which is overexpressed frequently in CRC tumours and is involved in the growth of CRC cells[22], and *SRC* (Fig. 3e), which encodes a non-receptor tyrosine kinase implicated in several human cancers including CRC[23].

Proteomic subtypes of CRC

The TCGA study reported three transcriptomic subtypes of CRC, designated 'microsatellite instability/CpG island methylator phenotype' (MSI/CIMP), 'invasive', and 'chromosomal instability' (CIN). Given the limited correlation between mRNA and protein levels, we asked whether CRC subtypes can be better represented with proteomics data. Using the consensus clustering[24] method (Methods and Extended Data Fig. 9), we identified five major proteomic subtypes in this tumour cohort, with 15, 9, 25, 11 and 19 cases in subtypes A to E, respectively (Fig. 4a).

We tested the association between the subtype classification and established genomic and epigenomic features of CRC using Fisher's exact test (Fig. 4b and Supplementary Table 11). Almost all hypermutated and MSI-high tumours were included in subtypes B and C, as well as tumours with *POLE* and *BRAF* mutations. However, statistically significant association with these features was only observed for subtype B (multiple-test adjusted $P < 0.05$). Moreover, subtype B was significantly associated with the TCGA CIMP -high methylation subtype, whereas subtype C was significantly associated with a non-CIMP subtype (cluster 4). Another unique feature of subtype B was the lack of *TP53* mutations and chromosome 18q loss. These results clearly established the association between proteomic subtype B and MSI-high and CIMP, but suggest that subtype C may have different biological underpinnings.

The remaining three subtypes were associated with CIN, another well-accepted genetic property of CRC. In particular, subtype E was significantly associated with both *TP53* mutations and 18q loss, genomic features frequently associated with CIN tumours[25]. Interestingly, subtype E was also associated with *HNF4A* amplification and relatively higher abundance of HNF4α protein (Fig. 4c). HNF4α abundance was significantly higher in subtype E tumours compared to normal colon samples (multiple-test adjusted $P = 1.09 \times 10^{-6}$, two-sided Wilcoxon rank-sum test); however, significant upregulation of HNF4α was not observed for other subtypes (Methods). This result, together with our reanalysis of shRNA knockdown data from the Achilles project (Extended Data Fig. 8), suggests that HNF4α dependency may be particularly associated with the subset of tumours or cells with *HNF4A* amplification.

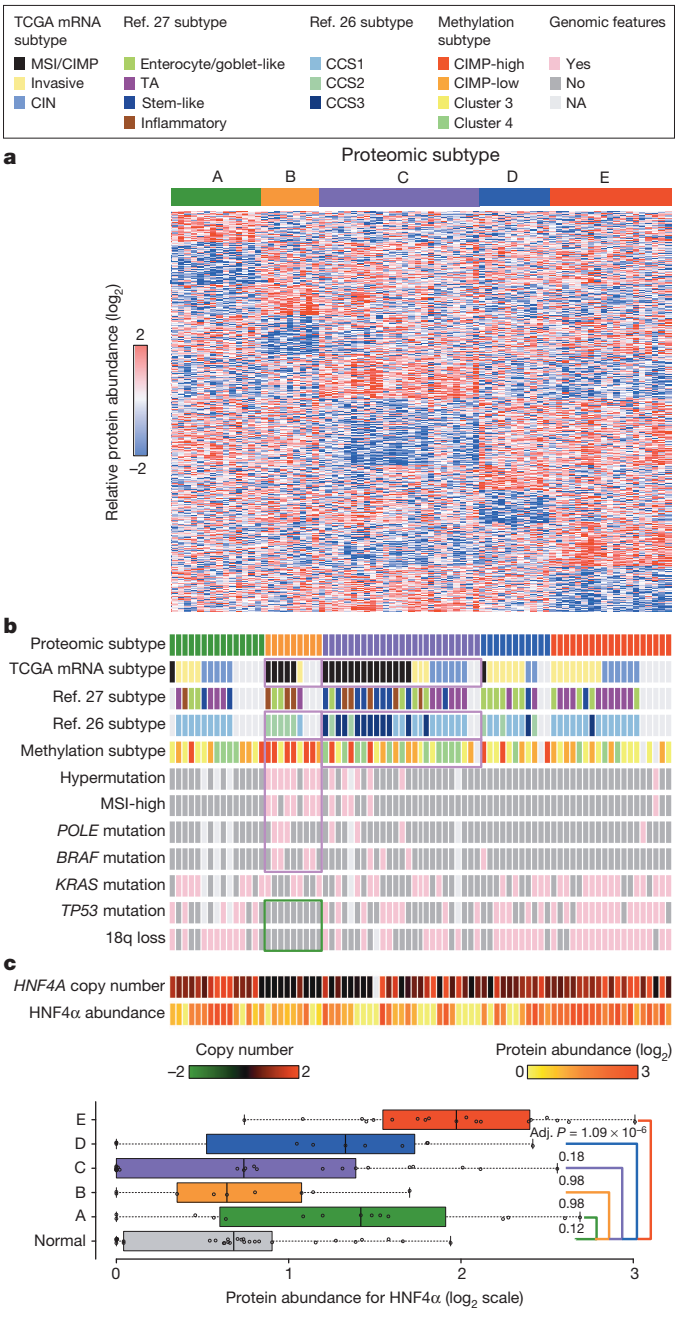


Figure 4 | Proteomic subtypes of colon and rectal cancers, associated genomic features, and relative abundance of HNF4α. **a**, Identification of five proteomic subtypes. Tumours are displayed as columns, grouped by proteomic subtypes as indicated by different colours. Proteins used for the subtype classification are displayed as rows. The heat map presents relative abundance of the proteins (logarithmic scale in base 2) in the 90-tumour cohort. **b**, Association of proteomic subtypes with major colorectal-cancer-associated genomic alterations and previously published transcriptomic and methylation subtypes. Subtypes that are significantly overlapped with a transcriptomic or methylation subtype are highlighted by pink boxes. Both proteomic subtypes B and C showed significant overlap with the TCGA MSI/CIMP subtype. In addition, they showed significant overlap with the CCS2 and CCS3 subtypes in the ref. 26 classification, respectively. Proteomic subtype B significantly overlapped with the TCGA CIMP-high methylation subtype, whereas subtype C significantly overlapped with a non-methylation subtype (TCGA cluster 4 methylation subtype). Subtypes overrepresented with a specific genomic alteration are also highlighted by pink boxes. The green box highlights the absence of *TP53* mutations and 18q loss in subtype B. **c**, The top panel shows *HNF4A* copy number and relative abundance of HNF4α protein in the five subtypes; the bottom panel compares relative abundance of HNF4α in the five subtypes to that in normal colon samples, respectively, and the adjusted *P* values are based on the two-sided Wilcoxon rank-sum test followed by multiple-test adjustment.

We also examined the association between the subtype classification and clinical features and found only that stage II tumours were significantly enriched in subtype C (multiple-test adjusted $P < 0.05$; Supplementary Table 11). Supervised statistical analyses at the individual protein level for 13 clinical and genomic features also identified few, if any, significant protein effects of these features, except for hypermutation status, MSI status and 18q loss (Supplementary Table 12), suggesting that the proteomic subtypes identified by the unsupervised clustering analysis captured the major proteome variations across the tumours.

Next, we compared the proteomic subtype classification with the TCGA transcriptional subtype classification for the 62 samples that had both subtype labels. Proteomic subtypes B and C both showed significant association with the TCGA subtype MSI/CIMP (Fig. 4b and Supplementary Table 11); however, they differ considerably at genomic, epigenomic and proteomic levels (Fig. 4a, b). We also examined alternative classifications of the TCGA samples based on two recently published transcriptomic subtype classifiers[26,27]. Proteomic subtype C, but not subtype B, showed enriched overlap with the 'stem-like' subtype described in ref. 27 and the colon cancer subtype 3 (CCS3) subtype described in ref. 26. Interestingly, tumours with stem-like and CCS3 classifications both have poor prognosis, which suggests that proteome subtype C also may be associated with poor prognosis. Therefore, the ability to distinguish subtype B from C through proteomics data are important, because MSI-high tumours typically have better prognosis[25].

Signatures for proteomic subtypes

To better understand the biology underlying the proteomic subtypes, we identified protein signatures for each subtype by supervised comparison of protein abundance in that subtype against all others; we also required signature proteins for a subtype to be significantly different in abundance compared to normal colon samples from 30 individuals analysed on the same proteome analysis platform (Methods and Supplementary Tables 13 and 14). As shown in Extended Data Fig. 10a, all CRC subtypes displayed more than 2,000 (>60%) significant protein abundance differences compared to normal colon. Although a full validation of the proteomic subtypes and protein signatures for the subtypes will require proteomic profiling data from an independent tumour cohort, a low cross-validation error rate of 3.8% demonstrated good generalizability of the subtypes and their signature proteins (Methods).

We performed Gene Ontology enrichment analysis for the subtype signatures using WebGestalt[28] (Methods and Supplementary Table 15). Genes involved in 'response to wounding' were significantly enriched in the up-signature of subtype C (multiple-test adjusted $P < 2.2 \times 10^{-16}$, Fisher's exact test). The wound-response gene signature is a powerful predictor of poor clinical outcome in patients with early stage breast cancers[29]. This result further links our subtype C to poor prognosis.

To understand better the functional networks underlying this subtype with potential clinical importance, we uploaded the up and down signatures of subtype C to NetGestalt[30] for enriched protein–protein interaction network module analysis. Four network modules were enriched with genes in the up signature for subtype C, whereas two modules were enriched with genes in the down signature (multiple-test adjusted $P < 0.01$, Fisher's exact test; Extended Data Fig. 10b). Notably, the down-signature-enriched module (III) included the E-cadherin (CDH1)-β-catenin (CTNNB1)-α-catenin (CTNNA1) complex (Extended Data Fig. 10c, e). E-cadherin, the most under-expressed protein in the subnetwork, suppresses invasion in lobular breast carcinoma[31] and is a switch for the epithelial-to-mesenchymal transition (EMT), which is associated with poor prognosis in colon cancer[32]. Other components of the module were desmosomal proteins (PKP2, JUP and DSG2) and cytokeratins (KRT18, KRT6A and KRT8). Reduction in both desmosome formation and cytokeratin expression is associated with EMT[33]. Moreover, proteins in the most significantly upregulated network module (Extended Data Fig. 10d, f) included collagens (COL1A1 and COL3A1) and extracellular matrix glycoproteins (FN1, BGN, FBN1 and FBN2) that also are markers of EMT[34,35]. These data strengthen the association of subtype C with poor prognosis and relate it to EMT activation.

Discussion

Our proteomic characterization of the genomically annotated TCGA colon tumours illustrates the power of integrated proteogenomic analysis. The data demonstrate that protein abundance cannot be reliably predicted from DNA- or RNA-level measurements. mRNA and protein levels were modestly correlated, as earlier cell and animal model studies suggested[36], but over two-thirds of these correlations were not statistically significant in the TCGA tumour set. Although most CNAs in CRC drive mRNA abundance changes, relatively few translated to consistent changes in protein abundance.

Genomic and proteomic technologies provide reinforcing data. RNA-seq data facilitate the discovery of variant proteins, which could serve as possible biomarker candidates or therapeutic targets. Combined mRNA and protein profiling data can identify potentially relevant genes in amplified chromosomal regions. This approach, which revealed the importance of chromosome 20q amplification and provided new insights into the role of HNF4α in CRC, can be broadly extended to understand roles of CNAs in other cancers. Proteomics identified CRC subtypes similar to those detectable by transcriptome profiles, but further captured features not detectable in transcript profiles. The separation of the TCGA MSI/CIMP subtype into distinct proteotypes illustrates the unique potential of proteomics-based subtyping. After validation in independent cohorts, protein subtype signatures could be directly translated into laboratory tests for tumour classification. Integrated proteogenomic analysis, as demonstrated in this study, will enable new advances in cancer biology, diagnostics and therapeutics.

METHODS SUMMARY

All tumour samples for the current study were obtained through the TCGA Biospecimen Core Resource (BCR) as described previously[6]. No other selection criteria other than availability were applied for this study. Patient-derived xenograft tumours from established basal and luminal B breast cancer intrinsic subtypes[37,38] were raised subcutaneously in 8-week-old NOD.Cg-Prkdcscid Il2rgtm1Wjl/SzJ mice (Jackson Laboratories, Bar Harbour, Maine) as described previously[39,40]. Normal colon biopsies were obtained from screening colonoscopies performed between July 2006 and October 2010 under Vanderbilt University Institutional Review Board (IRB) approval no. 061096.

Tissue proteins were extracted and tryptic peptide digests were analysed by multidimensional liquid chromatography-tandem mass spectrometry. Xenograft quality control samples were run after every five colorectal tumour samples. Raw data were processed for peptide identification by database and spectral library searching and identified peptides were assembled as proteins and mapped to gene identifiers for proteogenomic comparisons. Quantitative proteomic comparisons were based on spectral count data. Detailed descriptions of the samples, LC-MS/MS analysis, and data analysis methods can be found in the Methods. All of the primary mass spectrometry data on TCGA tumour samples are deposited at the CPTAC Data Coordinating Center as raw and mzML files and complete protein assembly data sets for public access (https://cptac-data-portal.georgetown.edu).

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 19 September 2013; accepted 2 May 2014.
Published online 20 July 2014.

1. The Cancer Genome Atlas Research Network Integrated genomic characterization of endometrial carcinoma. *Nature* **497,** 67–73 (2013).
2. The Cancer Genome Atlas Research Network Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455,** 1061–1068 (2008).
3. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* 474**,** 609–615 (2011); erratum **490,** 292 (2012).
4. The Cancer Genome Atlas Research Network Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489,** 519–525 (2012); corrigendum **491,** 288 (2012).
5. The Cancer Genome Atlas Research Network Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).

6. The Cancer Genome Atlas Research Network Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487,** 330–337 (2012).
7. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339,** 1546–1558 (2013).
8. Wang, X. & Zhang, B. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics* **29,** 3235–3237 (2013).
9. Wang, X. *et al.* Protein identification using customized protein sequence databases derived from RNA-Seq data. *J. Proteome Res.* **11,** 1009–1017 (2012).
10. Kim, W. K. *et al.* Identification and selective degradation of neopeptide-containing truncated mutant proteins in the tumors with high microsatellite instability. *Clin. Cancer Res.* **19,** 3369–3382 (2013).
11. Liu, H., Sadygov, R. G. & Yates, J. R. 3rd A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **76,** 4193–4201 (2004).
12. de Sousa Abreu, R., Penalva, L. O., Marcotte, E. M. & Vogel, C. Global signatures of protein and mRNA expression levels. *Mol. Biosyst.* **5,** 1512–1526 (2009).
13. Foss, E. J. *et al.* Genetic variation shapes protein networks mainly through non-transcriptional mechanisms. *PLoS Biol.* **9,** e1001144 (2011).
14. Ghazalpour, A. *et al.* Comparative analysis of proteome and transcriptome variation in mouse. *PLoS Genet.* **7,** e1001393 (2011).
15. Gry, M. *et al.* Correlations between RNA and protein expression profiles in 23 human cell lines. *BMC Genomics* **10,** 365 (2009).
16. Foss, E. J. *et al.* Genetic basis of proteome variation in yeast. *Nature Genet.* **39,** 1369–1375 (2007).
17. Fu, J. *et al.* System-wide molecular evidence for phenotypic buffering in *Arabidopsis. Nature Genet.* **41,** 166–167 (2009).
18. Peng, J. *et al.* Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Annals Applied Statistics* **4,** 53–77 (2010).
19. Garrison, W. D. *et al.* Hepatocyte nuclear factor 4α is essential for embryonic development of the mouse colon. *Gastroenterology* **130,** 19.e1–19.e (2006).
20. Chellappa, K., Robertson, G. R. & Sladek, F. M. HNF4α: a new biomarker in colon cancer? *Biomark. Med.* **6,** 297–300 (2012).
21. Cheung, H. W. *et al.* Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proc. Natl Acad. Sci. USA* **108,** 12372–12377 (2011).
22. Shimokawa, T. *et al.* Identification of TOMM34, which shows elevated expression in the majority of human colon cancers, as a novel drug target. *Int. J. Oncol.* **29,** 381–386 (2006).
23. Irby, R. B. *et al.* Activating *SRC* mutation in a subset of advanced human colon cancers. *Nature Genet.* **21,** 187–190 (1999).
24. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. R. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.* **52,** 91–118 (2003).
25. Fearon, E. R. Molecular genetics of colorectal cancer. *Annu. Rev. Pathol.* **6,** 479–507 (2011).
26. De Sousa E Melo, F. *et al.* Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nature Med.* **19,** 614–618 (2013).
27. Sadanandam, A. *et al.* A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nature Med.* **19,** 619–625 (2013).
28. Zhang, B., Kirov, S. & Snoddy, J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res.* **33,** W741–W748 (2005).
29. Chang, H. Y. *et al.* Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc. Natl Acad. Sci. USA* **102,** 3738–3743 (2005).
30. Shi, Z., Wang, J. & Zhang, B. NetGestalt: integrating multidimensional omics data over biological networks. *Nature Methods* **10,** 597–598 (2013).
31. Polyak, K. & Weinberg, R. A. Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits. *Nature Rev. Cancer* **9,** 265–273 (2009).
32. Loboda, A. *et al.* EMT is the dominant program in human colon cancer. *BMC Med. Genomics* **4,** 9 (2011).
33. Geiger, T., Sabanay, H., Kravchenko-Balasha, N., Geiger, B. & Levitzki, A. Anomalous features of EMT during keratinocyte transformation. *PLoS One* **3,** e1547 (2008).
34. Kiemer, A. K., Takeuchi, K. & Quinlan, M. P. Identification of genes involved in epithelial-mesenchymal transition and tumor progression. *Oncogene* **20,** 6679–6688 (2001).
35. Zeisberg, M. & Neilson, E. G. Biomarkers for epithelial-mesenchymal transitions. *J. Clin. Invest.* **119,** 1429–1437 (2009).
36. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Rev. Genet.* **13,** 227–232 (2012).
37. Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27,** 1160–1167 (2009).
38. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406,** 747–752 (2000).
39. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464,** 999–1005 (2010).
40. Li, S. *et al.* Endocrine-therapy-resistant *ESR1* variants revealed by genomic characterization of breast-cancer-derived xenografts. *Cell Rep.* **4,** 1116–1130 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by National Cancer Institute (NCI) CPTAC awards U24CA159988, U24CA160035, and U24CA160034; by NCI SPORE award P50CA095103 and NCI Cancer Center Support Grant P30CA068485; by National Institutes of Health grant GM088822; and by contract 13XS029 from Leidos Biomedical Research, Inc. Genomics data for this study were generated by The Cancer Genome Atlas pilot project established by the NCI and the National Human Genome Research Institute. Information about TCGA and the investigators and institutions comprising the TCGA research network can be found at http://cancergenome.nih.gov/.

Author Contributions B.Z., R.J.C.S., D.L.T., L.J.Z. and D.C.L. designed the proteomic analysis experiments, data analysis workflow, and proteomic–genomic data comparisons. K.F.S., L.J.Z., R.J.C.S. and D.C.L. directed and performed proteomic analysis of colon tumour and quality control samples. J.W., X.W., J.Z., Q.L., Z.S., P.W., S.W., R.J.C.S. and B.Z. performed proteomic-genomic data analyses. M.C.C., S.K., R.J.C.S. and D.L.T. performed analyses of mass spectrometry data and adapted algorithms and software for data analysis. S.R.D., R.R.T. and M.J.C.E. developed and prepared breast xenografts used as quality control samples. S.A.C., K.F.S. and D.C.L. designed strategy for quality control analyses. R.J.C.S., C.R.K, R.C.R. and H.R. coordinated acquisition, distribution and quality control evaluation of TCGA tumor samples. B.Z., J.W., R.J.C.S., R.J.C. and D.C.L. interpreted data in context of colon cancer biology. B.Z., R.J.C.S. and D.C.L. wrote the manuscript.

Author Information All of the primary mass spectrometry data on TCGA tumour samples are deposited at the CPTAC Data Coordinating Center as raw and mzML files for public access (https://cptac-data-portal.georgetown.edu). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.L. (daniel.liebler@vanderbilt.edu).

National Cancer Institute Clinical Proteomics Tumor Analysis Consortium (NCI CPTAC)

Steven A. Carr[1], Michael A. Gillette[1], Karl R. Klauser[1], Eric Kuhn[1], D. R. Mani[1], Philipp Mertins[1], Karen A. Ketchum[2], Amanda G. Paulovich[3], Jeffrey R. Whiteaker[3], Nathan J. Edwards[4], Peter B. McGarvey[4], Subha Madhavan[5], Pei Wang[6], Daniel Chan[7], Akhilesh Pandey[7], Ie-Ming Shih[7], Hui Zhang[7], Zhen Zhang[7], Heng Zhu[8], Gordon A. Whiteley[9], Steven J. Skates[10], Forest M. White[11], Douglas A. Levine[12], Emily S. Boja[13], Christopher R. Kinsinger[13], Tara Hiltke[13], Mehdi Mesri[13], Robert C. Rivers[13], Henry Rodriguez[13], Kenna M. Shaw[13], Stephen E. Stein[14], David Fenyo[15], Tao Liu[16], Jason E. McDermott[16], Samuel H. Payne[16], Karin D. Rodland[16], Richard D. Smith[16], Paul Rudnick[17], Michael Snyder[18], Yingming Zhao[19], Xian Chen[20], David F. Ransohoff[20], Andrew N. Hoofnagle[21], Daniel C. Liebler[22], Melinda E. Sanders[22], Zhiao Shi[22], Robbert J. C. Slebos[22], David L. Tabb[22], Bing Zhang[22], Lisa J. Zimmerman[22], Yue Wang[23], Sherri R. Davies[24], Li Ding[24], Matthew J. C. Ellis[24] & R. Reid Townsend[24]

[1]The Eli and Edythe L. Broad Institute of Massachusetts Institute of Technology and Harvard University Cambridge, Massachusetts 02142, USA. [2]Enterprise Science and Computing, Inc., 155 Gibbs St, Suite 420, Rockville, Maryland 20850, USA. [3]Clinical Research Division, Fred Hutchinson Cancer Research Center, 1100 Eastlake Avenue East, Seattle, Washington 98109, USA. [4]Department of Biochemistry and Molecular & Cellular Biology, Georgetown University Medical Center, 3900 Reservoir Rd NW, Washington, DC 20057, USA. [5]Innovation Center for Biomedical Informatics, Georgetown University Medical Center, 2115 Wisconsin Ave NW, Suite 110, Washington, DC 20057, USA. [6]Icahn Institute and Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, Hess CSM Building, Room S8-102, 1470 Madison Avenue, New York, New York 10029, USA. [7]Department of Pathology, The Johns Hopkins University, 600 North Wolfe Street, Baltimore, Maryland 21287, USA. [8]Department of Pharmacology and Molecular Science, the Johns Hopkins University, 733 N. Broadway, Baltimore, Maryland 21287, USA. [9]Antibody Characterization Laboratory, Advanced Technology Program, Leidos, Inc., 1050 Boyles Street, Frederick, Maryland 21701, USA. [10]Biostatistics Center, Massachusetts General Hospital Cancer Center, 55 Fruit Street, Boston, Massachusetts 02114, USA. [11]Department of Biological Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA. [12]Gynecology Service/Department of Surgery, Memorial Sloan Kettering Cancer Center, 1275 York Avenue, New York, New York 10065, USA. [13]Office of Cancer Clinical Proteomics Research, National Cancer Institute, 31 Center Drive, MS 2580 Bethesda, Maryland 20892, USA. [14]Biomolecular Measurement Division, Material Measurement Laboratory, National Institute of Standards and Technology, 100 Bureau Drive, M/S 8300, Gaithersburg, Maryland 20899, USA. [15]Department of Biochemistry and Molecular Pharmacology, Smilow Research Building, Room 201, 522 First Avenue, New York University Langone Medical Center, New York, New York 10016, USA. [16]Biological Sciences Division, Pacific Northwest National Laboratory, 902 Battelle Boulevard, Richland, Washington 99352, USA. [17]Spectragen-Informatics, Rockville, Maryland 20850, USA. [18]Department of Genetics, Stanford University, Stanford, California 94305, USA. [19]The Ben May Department for Cancer Research, University of Chicago, 929 East 57th Street, W421 Chicago, Illinois 60637, USA. [20]University of North Carolina at Chapel Hill, 130 Mason Farm Road, Chapel Hill, North Carolina 27599, USA. [21]Department of Lab Medicine, University of Washington, Campus Box 357110, Seattle, Washington 98195, USA. [22]Vanderbilt University School of Medicine, 1161 21st Avenue South, Nashville, Tennessee 37232, USA. [23]Bradley Department of Electrical and Computer Engineering, Virginia Tech, 900 N. Glebe Road, Arlington, Virginia 22203, USA. [24]Department of Medicine, Washington University in St. Louis, 660 S. Euclid Avenue, St. Louis, Missouri 063110, USA.

ARTICLE

doi:10.1038/nature13543

Molecular architecture and mechanism of the anaphase-promoting complex

Leifu Chang[1,2]†*, Ziguo Zhang[1,2]†*, Jing Yang[1,2]†, Stephen H. McLaughlin[2] & David Barford[1,2]†

The ubiquitination of cell cycle regulatory proteins by the anaphase-promoting complex/cyclosome (APC/C) controls sister chromatid segregation, cytokinesis and the establishment of the G1 phase of the cell cycle. The APC/C is an unusually large multimeric cullin–RING ligase. Its activity is strictly dependent on regulatory coactivator subunits that promote APC/C–substrate interactions and stimulate its catalytic reaction. Because the structures of many APC/C subunits and their organization within the assembly are unknown, the molecular basis for these processes is poorly understood. Here, from a cryo-electron microscopy reconstruction of a human APC/C–coactivator–substrate complex at 7.4 Å resolution, we have determined the complete secondary structural architecture of the complex. With this information we identified protein folds for structurally uncharacterized subunits, and the definitive location of all 20 APC/C subunits within the 1.2 MDa assembly. Comparison with apo APC/C shows that the coactivator promotes a profound allosteric transition involving displacement of the cullin–RING catalytic subunits relative to the degron-recognition module of coactivator and APC10. This transition is accompanied by increased flexibility of the cullin–RING subunits and enhanced affinity for UBCH10–ubiquitin, changes which may contribute to coactivator-mediated stimulation of APC/C E3 ligase activity.

Regulation of cell division by reversible protein phosphorylation and ubiquitination involves the coordinated interplay of protein kinases and phosphatases, and ubiquitin ligases and deubiquitinases[1]. The anaphase-promoting complex/cyclosome (APC/C) is an E3 cullin-RING ligase that mediates ubiquitin-dependent proteolysis of specific regulatory proteins to control chromosome segregation in mitosis, the events of cytokinesis and mitotic exit, maintenance of G1, and the initiation of DNA replication[2,3]. Through its cullin-RING catalytic module, the APC/C is related to the Skp1-Cullin-F-box protein (SCF) E3 ligases responsible for controlling S and G2 progression. However, compared to other RING ligases, the APC/C is unusually large. The reasons for its complexity likely arise from its role in orchestrating multiple cell cycle processes, its regulation by protein phosphorylation, and as the ultimate effector of the spindle assembly checkpoint[4].

The activity of the APC/C is tightly controlled by structurally related coactivator subunits (either CDC20 or CDH1). Their interaction with the APC/C is reciprocally regulated by CDK-dependent phosphorylation of core APC/C and coactivator subunits, resulting in a switch of CDC20 and CDH1 in late mitosis and change of substrate specificity. Coactivators recruit substrates to the APC/C by recognizing conserved destruction motifs or degrons—the D box[5] and KEN box[6]. In addition to their substrate recruitment roles, coactivators stimulate the E3 ligase activity of the APC/C through an unknown mechanism[7].

The activated APC/C-coactivator complex is an assembly of 15 different proteins in vertebrates (~1.2 MDa), of which only four have roles in catalysis (APC2 and APC11) and substrate recognition (APC10 and coactivator) (Extended Data Table 1a). Other large APC/C subunits function as molecular scaffolds to coordinate the juxtaposition of the catalytic and degron-recognition modules, although more direct roles in substrate and/or E2–ubiquitin interactions are also possible. Remarkably, all scaffolding subunits incorporate multiple repeat motifs. These include five TPR (tetratricopeptide repeat) proteins, the PC (proteasome/cyclosome) domain of APC1, and putative WD40 domain in APC4.

Reconstitution of recombinant APC/C facilitated its structural and biochemical analysis, providing definition of subunit stoichiometry and the location of many APC/C subunits within the electron microscopy-derived molecular envelope[8–10]. However, the structures and locations of numerous APC/C subunits remained unknown, and there were uncertainties regarding the location of its catalytic centre. To address these questions we determined a sub-nanometre resolution electron microscopy reconstruction of human APC/C as a ternary complex with the coactivator CDH1 and a high affinity substrate HSL1 (refs 11, 12) (APC/C–CDH1–HSL1). From this reconstruction we derived the complete secondary structure of the APC/C, allowing us to define the architectures and correctly assign positions of all APC/C subunits. In addition, we found that coactivator enhances APC/C affinity for UBCH10–ubiquitin, explaining at least in part how the coactivator stimulates APC/C ubiquitination activity. Comparison of ternary and apo APC/C structures defines the basis of the coactivator-mediated allosteric transition and provides insights into the mechanisms underlying this increased affinity for UBCH10–ubiquitin.

APC/C architecture

We determined a three dimensional reconstruction of the human APC/C–CDH1–HSL1 ternary complex to a nominal resolution of 7.4 Å (Fig. 1, Extended Data Figs 1 and 2 and Extended Data Table 1b). The resultant map is of sufficient quality to visualize α-helices as discrete rods and β-sheets as planar densities. Figure 1 shows three views of the APC/C with its molecular envelope colour-coded according to subunit assignments, together with the underlying secondary structure elements and architecture of individual APC/C subunits (Supplementary Video 1). Our higher resolution map indicates revised positions for APC1 and APC2. These subunits were previously assigned by means of antibody labelling[13,14] which guided docking of APC2 into a lower resolution electron microscopy map of budding yeast APC/C[15].

The APC/C adopts a triangular shape delineated by a lattice-like shell[8,13–17] (Fig. 1). The back and top of the complex is formed from the bowl-shaped

[1]Division of Structural Biology, Institute of Cancer Research, 237 Fulham Road, London SW3 6JB, UK. [2]MRC Laboratory of Molecular Biology, Cambridge CB2 0QH, UK. †Present address: MRC Laboratory of Molecular Biology, Cambridge CB2 0QH, UK (L.C., Z.Z., J.Y. and D.B.).
*These authors contributed equally to this work.

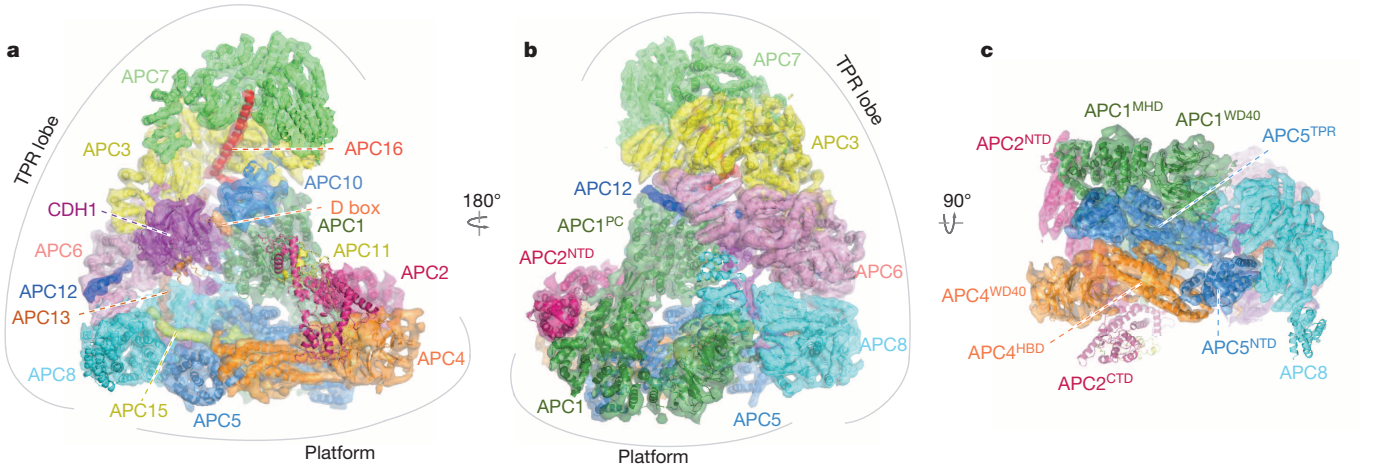


Figure 1 | Electron microscopy reconstructions of the *H. sapiens* APC/C–CDH1–HSL1 ternary complex at 7.4 Å resolution. **a**–**c**, Three views of the molecular envelope with the electron microscopy map segmented and colour-coded according to subunit assignments. The molecular envelope is represented as a transparent surface to view the underlying secondary structural elements and subunit architectures.

TPR lobe—an assembly of the four canonical TPR proteins and their accessory subunits (Extended Data Table 1a). The base of the APC/C comprises the platform subunits of APC4 and APC5, together with two domains of APC1. The carboxy-terminal PC domain of APC1 ($APC1^{PC}$) extends from the platform to contact the TPR lobe (Fig. 1b). The degron-recognition module of APC10 and CDH1 is located at the top of the cavity with APC10 interacting extensively with $APC1^{PC}$. The catalytic subunits APC2 (cullin) and APC11 (RING) of APC/C are positioned at the periphery of the platform such that the C-terminal domain of APC2 ($APC2^{CTD}$) and associated APC11 are at the front of the cavity below APC10 and CDH1 (Fig. 1a).

The TPR lobe is a quasi-symmetric array

The four evolutionarily related TPR proteins of the TPR lobe self-associate to form similar V-shaped homodimers (Fig. 2a–d). Each subunit comprises an α-helical solenoid with two turns of TPR helix[18–20]. With the amino-terminal TPR helix forming the homodimer interface, the C-terminal TPR helix creates a protein-binding groove. APC6 binds its accessory subunit APC12 through this groove (Fig. 2a)[18,19], whereas the APC3 and APC8 homodimers utilize one of their dyad-related C-terminal grooves to engage CDH1 (described below). APC10 contacts the TPR helical groove of the opposite subunit of the APC3 homodimer. Relative to the full-length *Schizosaccharomyces pombe* crystal structure[19], APC6 differs slightly in the context of the APC/C (Extended Data Fig. 3a and b and Supplementary Video 1).

TPR homodimers stack in parallel to generate a left-handed supra-helix featuring quasi-twofold symmetry (Figs 1, 2e and Supplementary Video 1)[20]. An ordered assembly of the TPR lobe is partly determined by the TPR accessory subunits APC13 and APC16, whose extended conformations span multiple TPR subunits (Fig. 2e). APC16 was identified as a long rod-like density and modelled as a 40-residue α-helix. It lies along the shallow grooves created by α-helices of TPR subunits APC3 and APC7 (interfaces i to iii) (Fig. 1 and Extended Data Fig. 3c). The location of APC16 agrees with its mapping to the upper region of the TPR lobe[21], and its co-purification with an APC3-APC7 sub-complex (data not shown). APC13 interacts with APC3, APC6 and APC8 (interfaces iv to vii) (Fig. 1 and Fig. 2e and Extended Data Fig. 3c), consistent with electron microscopy analysis of an APC13–green fluorescent protein fusion that positioned APC13 close to the APC6–APC3 interface[8].

Architectures of APC1, APC4 and APC5

APC4 and APC5 were assigned to the platform through antibody labelling[14], a negative stain electron microscopy analysis of an APC4–APC5 sub-complex[8], and assignment of toroidal density to the putative WD40 domain of APC4 (ref. 14) (Supplementary Video 1). Analysis of the platform in the APC/C–CDH1–HSL1 map reveals that the toroidal density comprises a WD40 domain ($APC4^{WD40}$) attached to a long helical bundle domain ($APC4^{HBD}$) (Fig. 3a and Extended Data Fig. 3d). This proposed architecture of APC4 is consistent with two-dimensional class views of an APC4–mFab complex derived using negatively stained electron micrographs (Extended Data Fig. 4), and with secondary structure predictions. Lying alongside $APC4^{HBD}$ is a TPR helix of 13 TPR motifs (Figs 1c and 3b) that corresponds to the TPR domain predicted for APC5 ($APC5^{TPR}$) (ref. 8). Connected to $APC5^{TPR}$, and at one end of $APC4^{HBD}$, is a globular

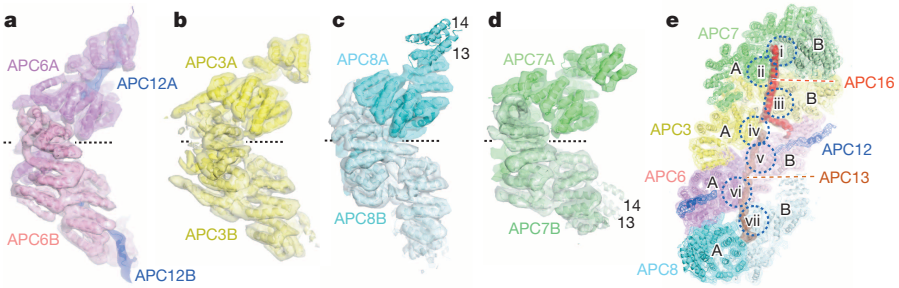


Figure 2 | The TPR homodimers assemble to form a quasi-symmetric array. **a**–**d**, The four canonical TPR homodimers form structurally related 'V'-shaped homodimers. Shown are the segmented electron microscopy maps and tertiary structures of APC6 (**a**), APC3 (**b**), APC8 (**c**) and APC7 (**d**). The two subunits of the homodimer are designated 'A' and 'B' and for APC6 are shown in two shades of magenta. C-terminal TPR motifs 13 and 14 of APC7B and APC8A project into solvent and are partially disordered. **e**, Front view of the TPR lobe showing how the TPR accessory subunits APC13 and APC16 span multiple TPR subunits. Roman numerals refer to interfaces indicated in Extended Data Fig. 3c.

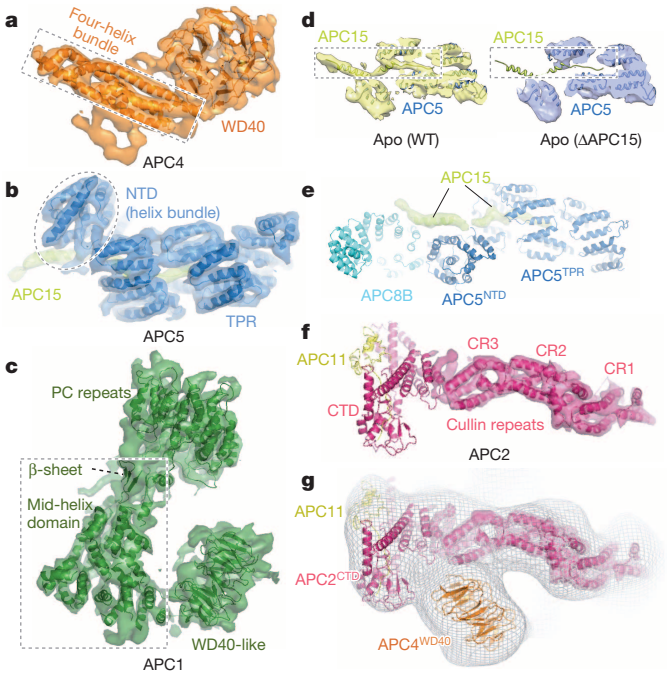


Figure 3 | Architecture of platform subunits. **a**, Overall view of APC4; **b**, APC5 with APC15; **c**, APC1. **d**, Deletion analysis reveals APC15 density in the vicinity of APC5: (left) apo APC/C, (right) apo APC/C$^{\Delta APC15}$. **e**, APC15-assigned density contacts APC8 and APC5. **f**, Only APC2^{NTD} (cullin repeats) is visible in the sharpened electron microscopy map, whereas the flexibility of the APC2^{CTD}–APC11 module results in weak fragmented density. **g**, An unsharpened map from three dimensional classification indicates the position of the APC2^{CTD}–APC11 module.

domain of nine α-helices (Figs 1c and 3b) that we assigned to the predicted N-terminal α-helical domain of APC5 (APC5^{NTD}).

APC1, unambiguously identified from its PC (proteasome-cyclosome) domain (APC1^{PC}), is located adjacent to APC2, APC5 and APC8 (Fig. 1a, b). APC1^{PC} resembles the PC domain of the proteasomal subunit Rpn2 (ref. 22) (Figs 1a, b, 3c and Extended Data Fig. 3e). Its eleven PC repeats assemble into a closed toroidal structure of two concentric rings of α-helices encircling two axial α-helices. Electron microscopy density corresponding to these 24 co-linear α-helices is resolved adjacent to APC10. The rod-shaped density immediately beneath APC1^{PC}, connecting it to the platform, resembles an α-helical solenoid (Figs 1b and 3c), and corresponds to the predicted ~300-residue mid-helical domain of APC1 (APC1^{MHD}). At the interface of the mid-helical and PC domains, a small β-sandwich structure (Fig. 3c) was assigned to the predicted β-sheet that immediately follows APC1^{PC}.

A third toroidal-shaped density, located adjacent to APC1^{MHD}, was assigned as the N-terminal WD40 domain of APC1 (APC1^{WD40}) (Fig. 3c and Extended Data Fig. 3f), consistent with the β-strand prediction for the N terminus of APC1. APC1^{WD40} tucks into the helical groove of APC5^{TPR} (Fig. 1c), and forms an edge-on contact with the C-terminal TPR helix of APC8 (Fig. 1b).

APC15 was identified from difference density from a reconstitution of the APC/C with APC15 deleted (APC/C$^{\Delta APC15}$) (Figs 1c, 3d and Extended Data Figs 1, 2). This revealed APC15 as the helical density bridging APC8 and APC5 (Fig. 3e), connected to an extended structure that inserts into the inner groove of the APC5^{TPR} superhelix (Fig. 3d, e). An interaction of APC15 with APC8 and platform subunits is in agreement with biochemical and structural evidence[8,9,23].

The catalytic module is flexible

For APC2, only α-helices of the N-terminal cullin repeats are clearly identified in the electron microscopy map (Fig. 1a and Fig. 3f). Density for APC2^{CTD} and APC11 is weak and fragmented. In an unsharpened map calculated to reveal low-resolution features, the APC2^{CTD}–APC11 module is recovered as a horn-like density (Fig. 3g), similar to a feature described in the 25 Å resolution cryo-electron microscopy map of apo *S. pombe* APC/C[16]. Three dimensional classification of the APC/C–CDH1–HSL1 electron microscopy data indicated structural variability of this density (Extended Data Fig. 5), consistent with the notion that the APC2^{CTD}–APC11 module is flexible. We used a well-defined three dimensional class to dock APC2^{CTD}–APC11, although in this density the APC11 RING domain cannot be located precisely (Fig. 3g). Because of the flexibility of APC2^{CTD}, we cannot exclude the possibility that APC11^{RING} is displaced from APC2^{CTD}, similar to the Rbx1 RING domain of the activated SCF[24,25].

Degron recognition by CDH1 and APC10

CDH1 and APC10 are in close proximity within the inner cavity adjacent to the APC3 homodimer[15,17] (Figs 1a, 4a). We fitted the β-sheets of the WD40 β-propeller domain of CDH1 (ref. 26) and the β-sandwich of APC10 (refs 27, 28) into their respective densities (Fig. 4a and Extended Data Fig. 3h, i). APC10 and CDH1 share structurally related C-terminal Ile-Arg motifs (IR tails) that interact with the C-terminal TPR motifs of APC3 (refs 12, 27, 29–31). Density for the IR tails of CDH1 and APC10 is visible extending from their globular domains (Fig. 4a). This shows that the two IR tails bind to two equivalent symmetry-related sites in the TPR helices of the APC3 homodimer (Fig. 4b). The three TPR motifs of the IR tail-binding sites were shown previously to be required for IR tail-dependent interactions of CDH1 to the APC/C[12].

Electron microscopy density for the HSL1 substrate is visible at the KEN box-recognition site centred on the upper surface of CDH1^{WD40} (ref. 26), whereas D box-assigned density is observed at the D box-coreceptor formed by CDH1 and APC10 (refs 15, 26, 30, 32–34) (Fig. 4a, d). Other segments of HSL1 are unstructured and not defined in the electron microscopy map. Crystallographic studies indicated that the D box-recognition site on coactivator is a channel on the edge of the WD40 domain that engages the D box motif RxxLxx(V/I/L) (residues P1 to P7), but not conserved polar residues P8 to P10 (ref. 26). In the APC/C–CDH1–HSL1 electron microscopy map, we observe tubular density that extends from the D box at the coactivator D box-binding site and connects with APC10. Modelling this density as P8 to P10 shows that these residues interact with a conserved polar surface on APC10 (Fig. 4d). Disruption of a loop that comprises this site (the 140s loop) impairs D box-dependent substrate recognition[32]. Ala substitutions of two hydrophilic residues (Ser 88 and Asn 147) in close proximity to P8 to P10 of the D box (Fig. 4d) attenuates APC/C^{CDH1} ubiquitination activity, as does mutation of D box residues P8 to P10 (Extended Data Fig. 6). Thus, the D box is a bipartite degron comprising a coactivator-interacting N-terminal (RxxLxx(V/I/L)) motif, and a hydrophilic C-terminal APC10-binding segment (Fig. 4e).

CDH1 promotes an allosteric transition

We could also visualize components of the N-terminal segment of CDH1 (CDH1^{NTD}). This forms a four α-helix bundle that packs co-linearly with the α-helices of APC1^{PC} and also contacts APC6 (Fig. 4f, g and Extended Data Fig. 3g). In addition, an extended segment of CDH1^{NTD} (L1) binds to the inner groove of the APC8B TPR helix, interacting with three TPR motifs that are structurally related to the IR tail-binding site of APC3 (Fig. 4c), and which contribute to coactivator binding[12,35].

The N terminus of coactivator stimulates APC/C ubiquitin ligase activity, independent of degron recognition[7]. To understand coactivator-mediated APC/C activation in the context of our complete subunit assignment, we determined three dimensional reconstructions of apo APC/C (Extended Data Figs 1 and 2). The most prominent structural difference between apo and ternary complexes involves a displacement of the platform subunits (Fig. 5a, b), similar to those described at lower resolution for *Xenopus* and human APC/C[13,14]. Moreover, in the apo state, electron microscopy density for APC2^{CTD} is more clearly defined (Fig. 5a, c) which, coupled to the lack of structural variability of APC2^{CTD}

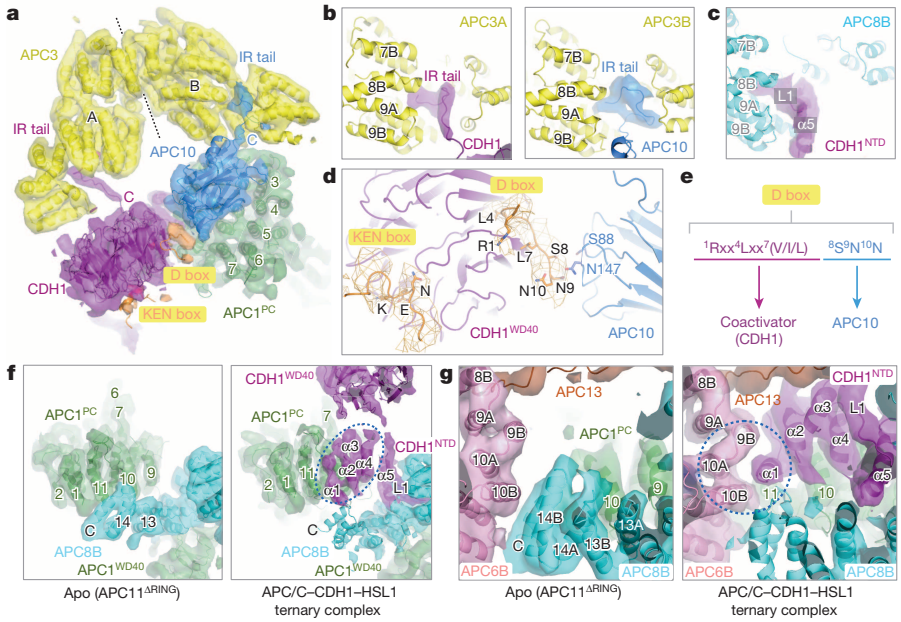


Figure 4 | Mechanism of D box recognition by CDH1 and APC10 and interactions with APC/C subunits. **a**, View showing IR tails of CDH1 and APC10 interacting with equivalent, symmetry-related sites on the APC3 homodimer. **b**, The IR tails of CDH1 and APC10 interact with equivalent TPR motifs 7 to 9 of symmetry-related subunits of the APC3 homodimer. **c**, CDH1^{NTD} interacts with APC8B, including TPR motifs equivalent to the IR tail-binding sites of APC3. **d**, Electron microscopy density (mesh) at the KEN box and D box sites of CDH1. Electron microscopy density for the D box interacting with two conserved loops of APC10 is visible. **e**, D box interacts as a bipartite degron with the D box coreceptor of CDH1 and APC10. Shown is a D box consensus sequence[26]. **f**, Interaction of CDH1^{NTD} with the APC1^{PC} displaces APC1^{PC}–APC8 interactions. **g**, CDH1^{NTD} interactions with APC6 (circled in right panel).

in three dimensional classes, indicates reduced structural flexibility of APC2^{CTD} (Extended Data Fig. 7).

Superimposing atomic models of apo and ternary states using APC1^{PC} and morphing between the two structures, defines the conformational transition induced on coactivator binding (Fig. 5b, d and Supplementary Video 2). This involves a rigid-body rotation of the platform and C-terminal TPR helix of APC8B around a hinge axis that bisects the APC1^{PC}–APC8B interface, propagating a conformational change of 20 Å to the peripheral APC2^{CTD}–APC11 catalytic module (Fig. 5 and Supplementary Video 2). Driving this rotation is the interaction of CDH1^{NTD} with APC1^{PC} and APC8B. In the apo state, the C-terminal TPR helix of APC8B is well ordered and interacts with APC1^{PC} (Fig. 4f). In the ternary complex, CDH1^{NTD} disrupts this interaction by binding to a site on APC1^{PC} that overlaps the site in contact with APC8B. This wedges APC1^{PC} and APC8B apart and disorders TPR motifs 13 and 14 of APC8B (Fig. 4f). A concomitant downwards-displacement of APC8B tilts the platform to translate the APC2^{CTD}–APC11 module upwards (Fig. 5d, e).

Coactivator-induced displacement and flexibility of the APC2^{CTD}–APC11 module has implications for understanding how the coactivator NTD stimulates APC/C ubiquitin ligase activity[7]. Increased catalytic activity could result from the flexibility of the APC2^{CTD}–APC11 module which, combined with its different location relative to the substrate-recognition module, may provide enhanced substrate (and ubiquitin polymer) access to the APC/C-bound E2–ubiquitin conjugate. Other possibilities include (1) enhancing the affinity of the APC/C for the E2–ubiquitin conjugate, (2) promoting the closed more reactive configuration of the E2–donor ubiquitin conjugate to stimulate efficient ubiquitin attachment to the target lysine[36–40], or (3) inducing an allosteric transition of the E2 (ref. 41).

To assess the effects of CDH1 on the affinity of APC/C for its cognate E2s we used surface plasmon resonance (SPR). Ternary APC/C bound UBCH10 with a K_D of 0.2 μM, an affinity some threefold higher than apo APC/C (Extended Data Fig. 9). Interestingly, coactivator promoted a more pronounced sevenfold increase in APC/C's affinity for UBCH10–Ub because ubiquitin conjugation of UBCH10 reduced its affinity for apo APC/C nearly threefold, whereas binding to ternary APC/C was only slightly affected (Extended Data Fig. 9). In contrast to UBCH10, ternary and apo APC/C had similar affinities for UBE2S (K_D 0.2 and 0.3 μM) (Extended Data Fig. 9), showing that coactivator-mediated APC/C activation does not enhance UBE2S affinity, and consistent with the idea that UBCH10 and UBE2S bind to distinct sites on the complex[40,42]. The C-terminal extension of UBE2S is required for synthesis of poly-ubiquitin chains on APC/C substrates[42–44]. Deleting this C terminus to yield the UBC domain alone (UBE2S-ΔC) abolished UBE2S binding to the core APC/C[44], a finding we confirmed for both apo and holo APC/C using SPR (Extended Data Fig. 9). The APC2^{CTD}–APC11 catalytic module had a low affinity for UBCH10 and did not interact with either UBCH10–Ub or UBE2S (Extended Data Fig. 9), indicating that high-affinity APC/C interactions with both UBCH10 and UBE2S depend on subunits other than APC2^{CTD}–APC11.

That ubiquitin modification of UBCH10 interferes with its binding to apo APC/C suggests that the energetically-favoured conformations of the UBCH10–Ub conjugate, including the catalytically-efficient closed conformation[36–40], are sterically hindered from binding apo APC/C. Thus, in addition to an increased affinity for UBCH10–Ub, a coactivator-dependent conformational change that allows association of UBCH10–Ub in the closed conformation, with enhanced intrinsic catalytic activity, could also contribute to stimulation of the E3 ligase activity of APC/C. It is therefore likely that, reminiscent of how Nedd8 conjugation stimulates the SCF[45], multiple mechanisms underlie coactivator-dependent stimulation of the APC/C, and potentially only UBCH10—responsible for initiating substrate ubiquitination[42,43,46]—is subject to coactivator-mediated regulation.

In conclusion, comparison of our APC/C–CDH1–HSL1 ternary structure with the electron microscopy reconstructions of APC/C^{MCC} (ref. 14) and APC/C–CDH1–EMI1 (ref. 47), determined at lower resolutions, shows that the presence of coactivator in all three complexes is associated with the same activated conformation of the platform subunits (Extended Data Fig. 8a, b). From this analysis it is clear that the MCC makes direct contact with the APC2^{CTD}–APC11 module, as does the

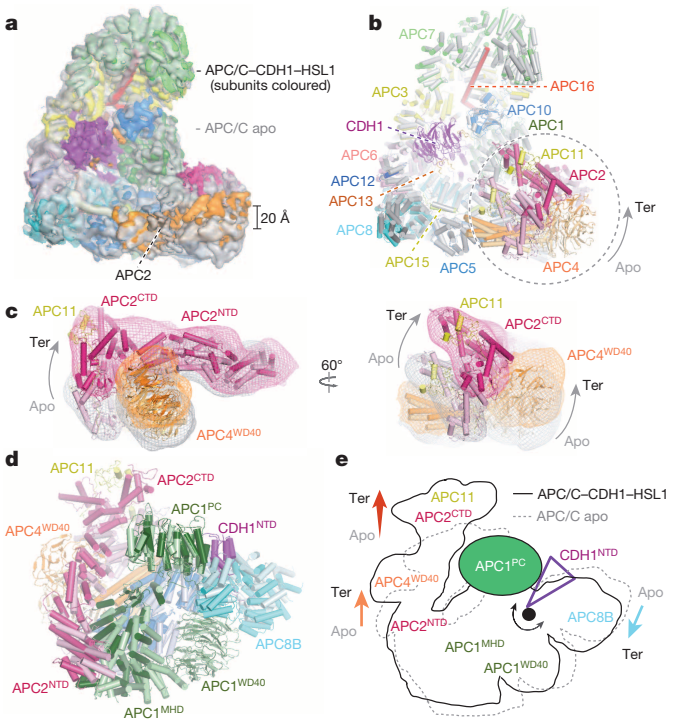


Figure 5 | Conformational change of the APC/C induced by CDH1^{NTD}. **a**, Superimposition of ternary and apo APC/C maps. The ternary map is colour-coded according to the subunits assignments of Fig. 1. Apo map is in grey. **b**, Superimposition of the apo and ternary coordinates. Apo coordinates are in grey except APC2, APC11 and APC4 shown in a lighter shade relative to the ternary complex. **c**, Two views showing the displacement of APC2–APC11 and APC4^{WD40}. **d**, View of rotation of the platform subunits on transition from apo to ternary states. Subunits of apo APC/C are shown in a lighter shade relative to the ternary complex. **e**, Schematic of the platform displacement. Insertion of CDH1^{NTD} (depicted as a wedge) between APC1^{PC} and APC8B causes platform rotation about the fixed APC1^{PC} domain. This propagates a 20-Å shift to APC2^{CTD}–APC11. The rotation axis is shown as a black circle.

density assigned to the inhibitory zinc-binding region (ZBR) and poly-basic tail of EMI1 (ref. 47), suggesting that these inhibitors may interfere with E2–Ub binding to APC2^{CTD}–APC11 and/or its flexibility.

A striking feature of the APC/C is the abundance of multiple repeat motif proteins (Extended Data Table 1a). The symmetry of TPR homodimers, coupled to the evolution of paralogues, may have resulted in the acquisition of replicated functions. For example, the IR tails of co-activator and APC10, and NTD of coactivator bind to equivalent sites on three TPR subunits (Fig. 4b, c). Our model of the APC/C provides insights into the assembly of other large multimeric complexes.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 29 October 2013; accepted 28 May 2014.
Published online 20 July 2014.

1. Teixeira, L. K. & Reed, S. I. Ubiquitin ligases and cell cycle control. *Annu. Rev. Biochem.* **82,** 387–414 (2013).
2. Pines, J. Cubism and the cell cycle: the many faces of the APC/C. *Nature Rev. Mol. Cell Biol.* **12,** 427–438 (2011).
3. Primorac, I. & Musacchio, A. Panta rhei: the APC/C at steady state. *J. Cell Biol.* **201,** 177–189 (2013).
4. Lara-Gonzalez, P., Westhorpe, F. G. & Taylor, S. S. The spindle assembly checkpoint. *Curr. Biol.* **22,** R966–R980 (2012).
5. Glotzer, M., Murray, A. W. & Kirschner, M. W. Cyclin is degraded by the ubiquitin pathway. *Nature* **349,** 132–138 (1991).
6. Pfleger, C. M. & Kirschner, M. W. The KEN box: an APC recognition signal distinct from the D box targeted by Cdh1. *Genes Dev.* **14,** 655–665 (2000).
7. Kimata, Y., Baxter, J. E., Fry, A. M. & Yamano, H. A role for the Fizzy/Cdc20 family of proteins in activation of the APC/C distinct from substrate recruitment. *Mol. Cell* **32,** 576–583 (2008).
8. Schreiber, A. *et al.* Structural basis for the subunit assembly of the anaphase-promoting complex. *Nature* **470,** 227–232 (2011).
9. Uzunova, K. *et al.* APC15 mediates CDC20 autoubiquitylation by APC/C(MCC) and disassembly of the mitotic checkpoint complex. *Nature Struct. Mol. Biol.* **19,** 1116–1123 (2012).
10. Zhang, Z. *et al.* Recombinant expression, reconstitution and structure of human anaphase-promoting complex (APC/C). *Biochem. J.* **449,** 365–371 (2013).
11. Burton, J. L. & Solomon, M. J. D box and KEN box motifs in budding yeast Hsl1p are required for APC-mediated degradation and direct binding to Cdc20p and Cdh1p. *Genes Dev.* **15,** 2381–2395 (2001).
12. Matyskiela, M. E. & Morgan, D. O. Analysis of activator-binding sites on the APC/C supports a cooperative substrate-binding mechanism. *Mol. Cell* **34,** 68–80 (2009).
13. Dube, P. *et al.* Localization of the coactivator Cdh1 and the cullin subunit Apc2 in a cryo-electron microscopy model of vertebrate APC/C. *Mol. Cell* **20,** 867–879 (2005).
14. Herzog, F. *et al.* Structure of the anaphase-promoting complex/cyclosome interacting with a mitotic checkpoint complex. *Science* **323,** 1477–1481 (2009).
15. da Fonseca, P. C. *et al.* Structures of APC/C(Cdh1) with substrates identify Cdh1 and Apc10 as the D-box co-receptor. *Nature* **470,** 274–278 (2011).
16. Ohi, M. D. *et al.* Structural organization of the anaphase-promoting complex bound to the mitotic activator Slp1. *Mol. Cell* **28,** 871–885 (2007).
17. Buschhorn, B. A. *et al.* Substrate binding on the APC/C occurs between the coactivator Cdh1 and the processivity factor Doc1. *Nature Struct. Mol. Biol.* **18,** 6–13 (2011).
18. Wang, J., Dye, B. T., Rajashankar, K. R., Kurinov, I. & Schulman, B. A. Insights into anaphase promoting complex TPR subdomain assembly from a CDC26–APC6 structure. *Nature Struct. Mol. Biol.* **16,** 987–989 (2009).
19. Zhang, Z., Kulkarni, K., Hanrahan, S. J., Thompson, A. J. & Barford, D. The APC/C subunit Cdc16/Cut9 is a contiguous tetratricopeptide repeat superhelix with a homo-dimer interface similar to Cdc27. *EMBO J.* **29,** 3733–3744 (2010).
20. Zhang, Z. *et al.* The four canonical TPR subunits of human APC/C form related homo-dimeric structures and stack in parallel to form a TPR suprahelix. *J. Mol. Biol.* **425,** 4236–4248 (2013).
21. Hutchins, J. R. *et al.* Systematic analysis of human protein complexes identifies chromosome segregation proteins. *Science* **328,** 593–599 (2010).
22. He, J. *et al.* The structure of the 26S proteasome subunit Rpn2 reveals its PC repeat domain as a closed toroid of two concentric alpha-helical rings. *Structure* **20,** 513–521 (2012).
23. Hall, M. C., Torres, M. P., Schroeder, G. K. & Borchers, C. H. Mnd2 and Swm1 are core subunits of the *Saccharomyces cerevisiae* anaphase-promoting complex. *J. Biol. Chem.* **278,** 16698–16705 (2003).
24. Duda, D. M. *et al.* Structural insights into NEDD8 activation of cullin-RING ligases: conformational control of conjugation. *Cell* **134,** 995–1006 (2008).
25. Calabrese, M. F. *et al.* A RING E3-substrate complex poised for ubiquitin-like protein transfer: structural insights into cullin-RING ligases. *Nature Struct. Mol. Biol.* **18,** 947–949 (2011).
26. He, J. *et al.* Insights into degron recognition by APC/C coactivators from the structure of an Acm1-Cdh1 complex. *Mol. Cell* **50,** 649–660 (2013).
27. Wendt, K. S. *et al.* Crystal structure of the APC10/DOC1 subunit of the human anaphase-promoting complex. *Nature Struct. Biol.* **8,** 784–788 (2001).
28. Au, S. W., Leng, X., Harper, J. W. & Barford, D. Implications for the ubiquitination reaction of the anaphase-promoting complex from the crystal structure of the Doc1/Apc10 subunit. *J. Mol. Biol.* **316,** 955–968 (2002).
29. Vodermaier, H. C., Gieffers, C., Maurer-Stroh, S., Eisenhaber, F. & Peters, J. M. TPR subunits of the anaphase-promoting complex mediate binding to the activator protein CDH1. *Curr. Biol.* **13,** 1459–1468 (2003).
30. Kraft, C., Vodermaier, H. C., Maurer-Stroh, S., Eisenhaber, F. & Peters, J. M. The WD40 propeller domain of Cdh1 functions as a destruction box receptor for APC/C substrates. *Mol. Cell* **18,** 543–553 (2005).
31. Thornton, B. R. *et al.* An architectural map of the anaphase-promoting complex. *Genes Dev.* **20,** 449–460 (2006).
32. Carroll, C. W., Enquist-Newman, M. & Morgan, D. O. The APC subunit Doc1 promotes recognition of the substrate destruction box. *Curr. Biol.* **15,** 11–18 (2005).
33. Chao, W. C., Kulkarni, K., Zhang, Z., Kong, E. H. & Barford, D. Structure of the mitotic checkpoint complex. *Nature* **484,** 208–213 (2012).
34. Tian, W. *et al.* Structural analysis of human Cdc20 supports multisite degron recognition by APC/C. *Proc. Natl Acad. Sci. USA* **109,** 18419–18424 (2012).
35. Izawa, D. & Pines, J. How APC/C-Cdc20 changes its substrate specificity in mitosis. *Nature Cell Biol.* **13,** 223–233 (2011).
36. Dou, H., Buetow, L., Sibbet, G. J., Cameron, K. & Huang, D. T. BIRC7–E2 ubiquitin conjugate structure reveals the mechanism of ubiquitin transfer by a RING dimer. *Nature Struct. Mol. Biol.* **19,** 876–883 (2012).
37. Plechanovová, A., Jaffray, E. G., Tatham, M. H., Naismith, J. H. & Hay, R. T. Structure of a RING E3 ligase and ubiquitin-loaded E2 primed for catalysis. *Nature* **489,** 115–120 (2012).
38. Pruneda, J. N. *et al.* Structure of an E3:E2~Ub complex reveals an allosteric mechanism shared among RING/U-box ligases. *Mol. Cell* **47,** 933–942 (2012).
39. Saha, A., Lewis, S., Kleiger, G., Kuhlman, B. & Deshaies, R. J. Essential role for ubiquitin-ubiquitin-conjugating enzyme interaction in ubiquitin discharge from Cdc34 to substrate. *Mol. Cell* **42,** 75–83 (2011).
40. Wickliffe, K. E., Lorenz, S., Wemmer, D. E., Kuriyan, J. & Rape, M. The mechanism of linkage-specific ubiquitin chain elongation by a single-subunit e2. *Cell* **144,** 769–781 (2011).

41. Das, R. *et al.* Allosteric activation of E2-RING finger-mediated ubiquitylation by a structurally defined specific E2-binding region of gp78. *Mol. Cell* **34,** 674–685 (2009).
42. Williamson, A. *et al.* Identification of a physiological E2 module for the human anaphase-promoting complex. *Proc. Natl Acad. Sci. USA* **106,** 18213–18218 (2009).
43. Wu, T. *et al.* UBE2S drives elongation of K11-linked ubiquitin chains by the anaphase-promoting complex. *Proc. Natl Acad. Sci. USA* **107,** 1355–1360 (2010).
44. Wang, W. & Kirschner, M. W. Emi1 preferentially inhibits ubiquitin chain elongation by the anaphase-promoting complex. *Nature Cell Biol.* **15,** 797–806 (2013).
45. Saha, A. & Deshaies, R. J. Multimodal activation of the ubiquitin ligase SCF by Nedd8 conjugation. *Mol. Cell* **32,** 21–31 (2008).
46. Jin, L., Williamson, A., Banerjee, S., Philipp, I. & Rape, M. Mechanism of ubiquitin-chain formation by the human anaphase-promoting complex. *Cell* **133,** 653–665 (2008).
47. Frye, J. J. *et al.* Electron microscopy structure of human APC/C(CDH1)-EMI1 reveals multimodal mechanism of E3 ligase shutdown. *Nature Struct. Mol. Biol.* **20,** 827–835 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank P. da Fonseca for her guidance preparing cryo electron microscopy grids and X. Bai and S. Scheres for help with the use of RELION and A. Plechanovova for advice in preparing stable UBCH10–Ub conjugates. We thank D. Morgan and W. Chao for their comments on the manuscript and D. Morgan for communicating data before publication. This work was funded by a Cancer Research UK grant to D.B.

Author Contributions L.C. prepared grids, collected and analysed electron microscopy data and determined the three dimensional reconstructions, fitted coordinates and built models, prepared figures and co-wrote the paper. Z.Z. designed and made constructs, performed biochemical analysis and purified proteins. J.Y. prepared and purified the complexes. S.H.McL. performed and analysed SPR experiments. D.B. directed the project, built models and co-wrote the paper.

Author Information Electron microscopy maps are deposited with the EMDataBank with accession codes: EMD-2651 (ternary), EMD-2652 (apo), EMD-2653 (APC/C–APC11$^{\Delta RING}$) and EMD-2654 (APC/C$^{\Delta APC15}$). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.B. (dbarford@mrc-lmb.cam.ac.uk).

LETTER

doi:10.1038/nature13616

A massive galaxy in its core formation phase three billion years after the Big Bang

Erica Nelson[1], Pieter van Dokkum[1], Marijn Franx[2], Gabriel Brammer[3], Ivelina Momcheva[1], Natascha Förster Schreiber[4], Elisabete da Cunha[5], Linda Tacconi[4], Rachel Bezanson[6], Allison Kirkpatrick[7], Joel Leja[1], Hans-Walter Rix[5], Rosalind Skelton[8], Arjen van der Wel[5], Katherine Whitaker[9] & Stijn Wuyts[4]

Most massive galaxies are thought to have formed their dense stellar cores in early cosmic epochs[1–3]. Previous studies have found galaxies with high gas velocity dispersions[4] or small apparent sizes[5–7], but so far no objects have been identified with both the stellar structure and the gas dynamics of a forming core. Here we report a candidate core in the process of formation 11 billion years ago, at redshift $z = 2.3$. This galaxy, GOODS-N-774, has a stellar mass of 100 billion solar masses, a half-light radius of 1.0 kiloparsecs and a star formation rate of 90^{+45}_{-20} solar masses per year. The star-forming gas has a velocity dispersion of 317 ± 30 kilometres per second. This is similar to the stellar velocity dispersions of the putative descendants of GOODS-N-774, which are compact quiescent galaxies at $z \approx 2$ (refs 8–11) and giant elliptical galaxies in the nearby Universe. Galaxies such as GOODS-N-774 seem to be rare; however, from the star formation rate and size of this galaxy we infer that many star-forming cores may be heavily obscured, and could be missed in optical and near-infrared surveys.

We identified the candidate forming core, GOODS-N-774, using the 3D-HST catalogues in the five CANDELS (Cosmic Assembly Near-Infrared Deep Extragalactic Legacy Survey) fields[12]. GOODS-N-774 has a circularized effective radius of $r_e = 1.0$ kpc from Hubble Space Telescope (HST) F160W (H_{160}) Wide Field Camera 3 (WFC3) imaging;[13] a stellar mass of $1.0 \times 10^{11} M_\odot$ (refs 12, 14; $M_\odot$, solar mass); rest-frame UVJ colours consistent with a star-forming galaxy; and a Spitzer Multiband Imaging Photometer (MIPS) 24 μm flux of 104 μJy. Figure 1 shows the stellar mass density profile derived from the observed H_{160} surface brightness profile corrected for the HST point spread function[15]. The surface density profile is strikingly similar to the average profile of massive quiescent galaxies at $z \approx 2$ (red line), and much more concentrated than the average profile of massive star-forming galaxies at that redshift[13] (light blue).

The near-infrared spectrum of GOODS-N-774 is shown in Fig. 2. The continuum is clearly detected, along with emission lines that we identify as Hα and [N II] redshifted to $z = 2.300$. The gas velocity dispersion is $\sigma = 317 \pm 30$ km s^{-1}, equivalent to a full-width at half-maximum of 750 km s^{-1}. Typically, objects with such large linewidths are mergers or are dominated by active galactic nuclei[4] (AGNs). If the line emission in GOODS-N-774 is partly or largely due to the presence of an AGN, its velocity dispersion, size and stellar mass measurements would not be reliable.

There is no evidence for the presence of an AGN in GOODS-N-774. It is not detected in the deep Chandra 2 Ms X-ray data in GOODS-North with an X-ray luminosity upper limit of $L_X < 1.2 \times 10^{42}$ erg s^{-1}. Although an AGN cannot be conclusively ruled out, this upper limit is consistent with the star formation rate of the galaxy. Also, the galaxy has line ratios [O III]/[O II] = 0.7 ± 0.5, [O III]/Hβ = 1.2 ± 0.9 and [N II]/Hα = 0.4 ± 0.1, indicating that the gas is in a low-ionization state. Therefore, stellar photoionization and, hence, ultimately, star formation, is the likely origin of the line emission. Finally, the observed infrared spectral energy distribution (SED) requires strong polycyclic aromatic hydrocarbon emission to simultaneously explain the MIPS 24 μm and Herschel data (Fig. 3), effectively ruling out the presence of a dominant AGN. We quantified this by fitting composite SEDs with varying AGN

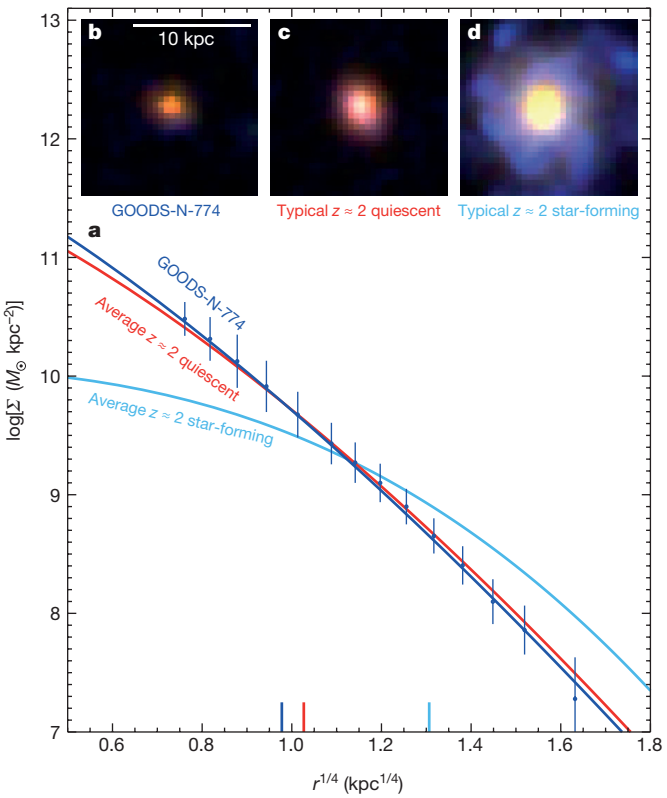


Figure 1 | Structural properties of GOODS-N-774. a, Surface density profile of GOODS-N-774 (blue line), as derived from deep WFC3 H_{160} imaging. Error bars are s.d. The galaxy has a mass of $1.0 \times 10^{11} M_\odot$ and an effective radius of $r_e = 1.0$ kpc. The light blue curve shows the average profile of 67 star-forming galaxies at $1.9 < z < 2.1$ with $10.9 < \log(M_{stellar}) < 11.2$ (refs 12, 13). The red curve shows the average profile of 24 quiescent galaxies with the same mass and redshift selection criteria. **b–d**, Colour images show GOODS-N-774 (**b**), a typical quiescent galaxy (**c**) and a typical star-forming galaxy (**d**). Vertical bars along the *x* axis in **a** indicate effective radii for the galaxies in **b–d** (colour-coded). The structure of GOODS-N-774 is similar to that of massive quiescent galaxies.

[1]Astronomy Department, Yale University, New Haven, Connecticut 06511, USA. [2]Leiden Observatory, Leiden University, NL-2300 RA Leiden, The Netherlands. [3]Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, Maryland 21218, USA. [4]Max-Planck-Institut für Extraterrestrische Physik, Giessenbackstrasse 1, 85748 Garching, Germany. [5]Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany. [6]Steward Observatory, University of Arizona, 933 North Cherry Avenue, Tucson, Arizona 85721, USA. [7]Department of Astronomy, University of Massachusetts, Amherst, Massachusetts 01002, USA. [8]South African Astronomical Observatory, PO Box 9, Observatory, Cape Town 7935, South Africa. [9]Astrophysics Science Division, Goddard Space Center, Greenbelt, Maryland 20771, USA.

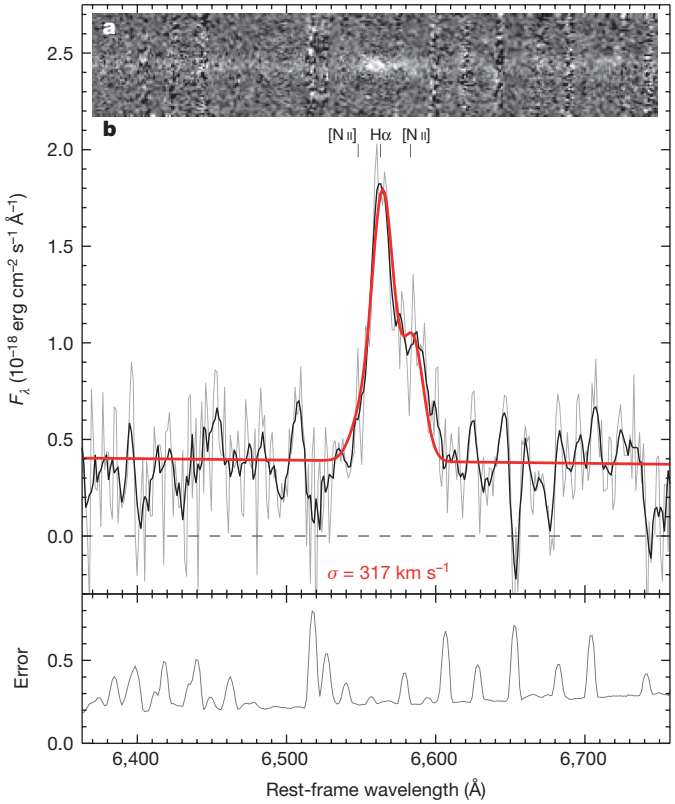


Figure 2 | Velocity dispersion of GOODS-N-774. Near-infrared spectrum in two dimensions (**a**) and 1D (**b**) obtained with NIRSPEC on Keck I. The grey curve is at the original resolution; the black curve shows the spectrum smoothed with a 20 Å boxcar. The best-fit Gaussians to the Hα $\lambda = 6{,}563$ Å and [N II] $\lambda = 6{,}548$ and 6,584 Å emission lines are shown in red. The velocity dispersion is 317 ± 30 km s^{-1}, which is equivalent to an inclination-corrected circular velocity of $V_{\rm rot} \approx 680$ km s^{-1} if the gas is rotating in a disk. The rest-frame equivalent width of Hα is 66 ± 8 Å and its luminosity is $(3.4 \pm 0.4) \times 10^{42}$ erg s^{-1}.

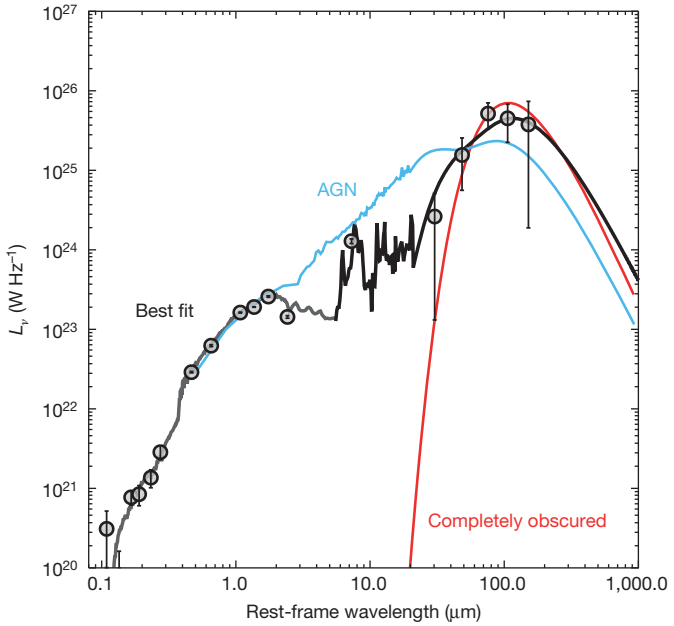


Figure 3 | Ultraviolet/far-infrared spectral energy distribution of GOODS-N-774. Rest-frame ultraviolet/far-infrared photometry of GOODS-N-774. Error bars are s.d. A stellar population synthesis model fit[14] to the ultraviolet/near-infrared SED is shown in grey. The black line shows the composite star-formation/AGN SED[16] that is the best fit to the mid- and far-infrared data. This best fit has no AGN contribution and implies a star formation rate of $90 M_\odot$ yr^{-1}. For reference, the light blue line shows a composite SED with an AGN contribution of 80%. The red curve shows a black body with a size of 1 kpc and a bolometric luminosity of 10^{12} solar luminosities.

contributions[16]. The best fit is obtained for a pure star-forming template with no AGN contribution (Fig. 3).

We infer that the linewidth of GOODS-N-774 is among the highest measured for a normal star-forming galaxy at high redshift (Extended Data Fig. 1). If the gas is in a disk, it is rotating with a velocity of $V_{\rm rot} \approx 550$ km s^{-1}, or $V_{\rm rot} \approx 680$ km s^{-1} after correcting for inclination. The observed gas velocity dispersion of 317 km s^{-1} is similar to the median stellar velocity dispersion of 304 km s^{-1} in a sample of quiescent galaxies at $z = 1.5$–2.2 with median mass $1.9 \times 10^{11} M_\odot$ (refs 8–11; Fig. 4).

The inferred dynamical mass is $1.5 \times 10^{11} M_\odot$, which is roughly 1.5 times the stellar mass, suggesting a gas fraction of $\lesssim 50\%$. In Fig. 4, we explicitly compare the dynamical and structural properties of GOODS-N-774 with those of galaxies in the Sloan Digital Sky Survey (SDSS) and those of quiescent galaxies at $z \approx 2$. The galaxy has a much smaller size and a higher velocity dispersion than do SDSS galaxies of the same total dynamical mass. Its properties are very similar to those of the samples of quiescent galaxies at $z \approx 2$ that have been compiled over the past few years, and we infer that we have identified an example of star-forming galaxies in this region of parameter space.

The Hα luminosity is $(3.4 \pm 0.4) \times 10^{42}$ erg s^{-1}, which implies a minimum star formation rate (with no reddening correction) of $\sim 16 M_\odot$ yr^{-1} for a Chabrier initial mass function[17,18]. The red colour of the galaxy ($R_{606} - H_{160} = 4.2$ mag), and the fact that it is detected with MIPS and Herschel, suggests that the actual, dust-corrected star formation rate is much higher. The 24 μm flux alone indicates a star formation rate of $135 M_\odot$ yr^{-1} (ref. 19). Fitting the 24 μm–500 μm data with empirical composite star-forming SEDs[16] or theoretical models[20] gives slightly lower values than do the 24 μm data alone, and we infer that the star formation rate is 90^{+45}_{-20} $M_\odot$ yr^{-1}. This confirms that the star formation is highly obscured, with ~3 mag of extinction in the direction of the Hα emission and an infrared/ultraviolet luminosity ratio of $\gtrsim 200$.

GOODS-N-774 has a specific star formation rate of $\sim 1 \times 10^{-9}$ yr^{-1}, which places it on the star-forming sequence at $z = 2.3$ (ref. 19). If the galaxy had a constant star formation rate leading up to the epoch of observation, then its mass was built up over a period of ~1 Gyr since $z \approx 3.3$. Although short compared with the age of the Universe at $z = 2.3$, this build-up phase is ~200 dynamical times, which is longer than expected from the Kennicutt–Schmidt law[21]. This suggests that the galaxy had a higher star formation rate in the past, or that the star formation rate has been throttled by the rate of gas accretion onto the halo: a galaxy with a stellar mass of $M_* = 1.0 \times 10^{11} M_\odot$ would have a baryonic accretion rate of $\sim (60 - 120) M_\odot$ yr^{-1} (ref. 22), in good agreement with the observed star formation rate.

The gas in a galaxy such as this, growing by means of rapid star formation in a deep gravitational potential well, should be rapidly enriched with metals, and we would thus expect it to exhibit a high gas-phase metallicity. This is consistent with what we observe: the galaxy has [N II]/Hα $= 0.4 \pm 0.1$, which implies a high metallicity ($12 + \log({\rm O/H}) \approx 9.05$, although the conversion[23] is somewhat uncertain). After the star formation phase, the gas is probably heated, expelled or both[2,22]. The quiescent core that remains will then probably evolve into a giant elliptical galaxy[2,3] with a central stellar metallicity that is similar to the gas-phase metallicity of the star-forming core at high redshift[24].

Galaxies such as GOODS-N-774 are rare. Candidate compact star-forming galaxies with less extreme properties have been identified in fairly large numbers[5,6], but in the ~900 arcmin^2 of the five 3D-HST CANDELS fields there are only three objects at $2 < z < 3$ with 24 μm fluxes of ≥100 μJy, high central mass densities ($\log(M/M_\odot)_{r<1\,{\rm kpc}} \geq 10.5$; ref. 7) and concentrated stellar distributions ($r_{\rm e} \leq 1$ kpc). We observed all three galaxies with Keck I, and GOODS-N-774 is the only

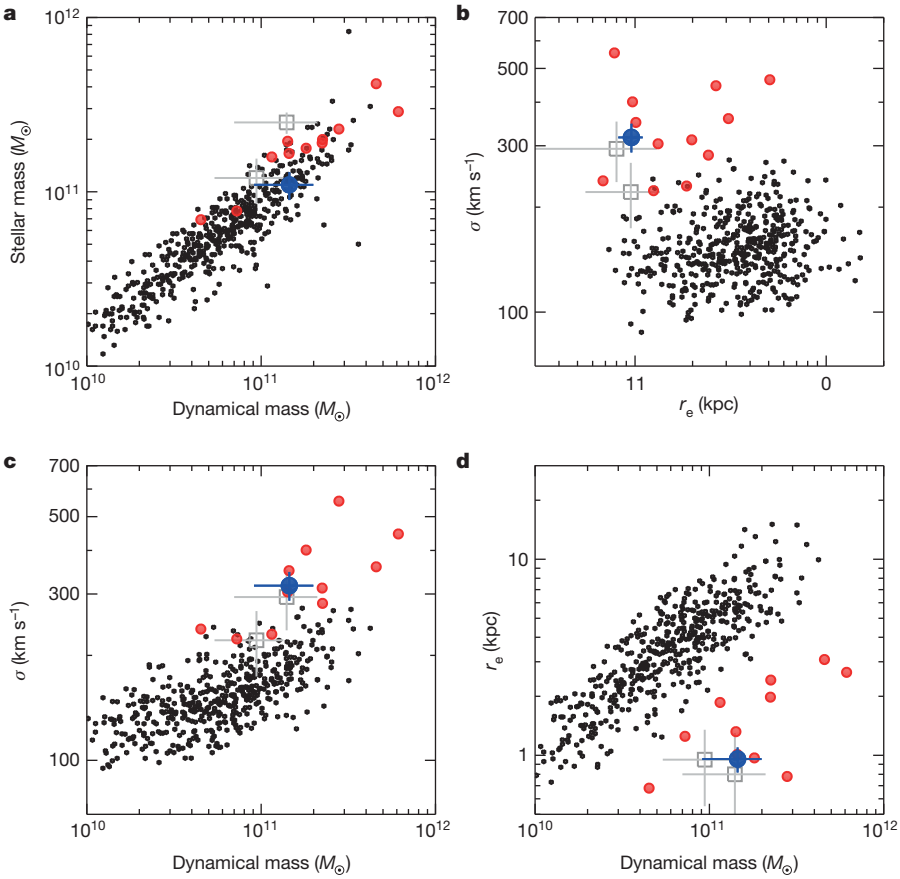


Figure 4 | Properties of GOODS-N-774 compared with quiescent galaxies. Comparisons of the size, mass and gas dynamics of GOODS-N-774 (blue symbols) to the sizes, masses and stellar dynamics of galaxies in the SDSS (black) and massive quiescent galaxies at $z \approx 2$ (red)[8–11]. GOODS-N-774 has properties that are similar to previously studied massive quiescent galaxies at $z \approx 2$ and is substantially offset from nearby galaxies. CO dynamics and CO sizes of two compact SMGs from ref. 13 (HDF 76 and N2850.2) are shown in grey. Error bars are s.d.

confirmed candidate: GOODS-S-5981[6] has a narrow linewidth, whereas COSMOS-8388 is difficult to interpret because it has an AGN. The number density we infer is $\sim 10^{-6}$ Mpc^{-3} (including all three candidates), compared with $\sim 10^{-4}$ Mpc^{-3} for the overall population of galaxies with dense cores at $z \approx 2$ (ref. 3).

This mismatch could imply that the lifetime of the compact star-forming phase is very short, as has been suggested previously on the basis of similar number density arguments[4]. It may be that we are witnessing the aftermath of the contraction of a gravitationally unstable star-forming disk[25] or of a merger of large star-forming galaxies[4]. However, neither tidal features nor extended wings are apparent in the surface density distribution.

It is perhaps more likely that the lifetime of the compact star-forming phase is relatively long and that we are missing many star-forming compact galaxies in present surveys. From the compact morphology and high star formation rate, we infer a high gas column density for this object[21]: $N_H = 2.6 \times 10^{23}$ cm^{-2}. This gas column density is nearly an order of magnitude higher than in an average ultraviolet-selected star-forming galaxy at the same cosmic epoch[26], and 2.5 orders of magnitude higher than in the disk of a typical galaxy in the local Universe[21]. This high column density of gas in conjunction with the abundance of metals implies[27] a very high extinction: $A_V \gtrsim 100$ mag for a screen (V, visual band), and $A_V \gtrsim 6$ mag if the dust and the stars are mixed. The amount of extinction is driven by the dust column density, not the dust mass, meaning that at fixed dust mass, a compact galaxy will be more obscured than a large galaxy. The detection of rest-frame optical flux, and of Hα emission, is inconsistent with such high values for extinction. The dust distribution is probably non-uniform, and it may be that, for GOODS-N-774, we are looking along a relatively unobscured line of sight.

More typical star-forming cores could be entirely obscured[27,28], and begin to resemble black bodies with temperatures of $\sim$30 K (red curve in Fig. 3; calculated using a radius of 1 kpc and a bolometric luminosity of 10^{12} solar luminosities). It may be possible to select such obscured progenitors at long wavelengths, near the peak of the redshifted dust emission. It has been demonstrated that redshifts, sizes and velocity widths of infrared-luminous galaxies can be measured from CO emission. In fact, the closest analogues to GOODS-N-774 are the two submillimetre-selected galaxies (SMGs) HDF 76 and N2850.2 (Fig. 4), which have high linewidths and small sizes in the CO line[4]. It will be interesting to determine whether the stellar distributions of these galaxies are similar to the gas distribution, or whether these are dense star-forming regions inside larger galaxies.

Longer-wavelength studies of large, unbiased samples can show whether GOODS-N-774 is an example of a parent population of compact star-forming galaxies that are heavily obscured[4]. There may also be multiple paths to a compact, quiescent galaxy: some (such as HDF 76 and N2850.2[4]) may form most of their stars in mergers with star formation rates of $\gtrsim 1{,}000 M_\odot$ yr^{-1}, whereas others (such as GOODS-N-774) may grow relatively slowly in an obscured, accretion-throttled mode. Whatever the dominant mode turns out to be, because the stars in dense cores account for 10–20% of the total $z \approx 2$ stellar mass density[3], star-forming cores should account for a significant fraction of all star formation in the high-redshift Universe.

Very recently, evidence supporting our conclusions has been posted online[29].

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 4 May; accepted 19 June 2014.
Published online 27 August 2014.

1. Daddi, E. *et al.* Passively evolving early-type galaxies at $1.4 \lesssim z \lesssim 2.5$ in the Hubble Ultra Deep Field. *Astrophys. J.* **626,** 680–697 (2005).
2. Oser, L., Ostriker, J. P., Naab, T., Johansson, P. H. & Burkert, A. The two phases of galaxy formation. *Astrophys. J.* **725,** 2312–2323 (2010).
3. van Dokkum, P. *et al.* Dense cores in galaxies out to $z = 2.5$ in SDSS, UltraVISTA, and the five 3D-HST/CANDELS fields: number density, evolution, and the

apparent need for efficient cooling at high redshift. *Astrophys. J.* (submitted); preprint at http://arxiv.org/abs/1404.4874 (2014).
4. Tacconi, L. *et al.* Submillimeter galaxies at $z \sim 2$: evidence for major mergers and constraints on lifetimes, IMF, and CO-H_2 conversion factor. *Astrophys. J.* **680,** 246–262 (2008).
5. Toft, S. *et al.* Submillimeter galaxies as progenitors of compact quiescent galaxies. *Astrophys. J.* **782,** 68 (2014).
6. Barro, G. *et al.* CANDELS+3D-HST: compact SFGs at $z \sim 2-3$, the progenitors of the first quiescent galaxies. Preprint at http://arxiv.org/abs/1311.5559 (2013).
7. Williams, C. C. *et al.* The progenitors of the compact early-type galaxies at high redshift. *Astrophys. J.* **780,** 1 (2014).
8. Bezanson, R., van Dokkum, P., van de Sande, J., Franx, M. & Kriek, M. Massive and newly dead: discovery of a significant population of galaxies with high-velocity dispersions and strong Balmer lines at $z \sim 1.5$ from deep Keck spectra and HST/WFC3 imaging. *Astrophys. J.* **764,** L8 (2013).
9. van Dokkum, P. G., Kriek, M. & Franx, M. A high stellar velocity dispersion for a compact massive galaxy at redshift $z = 2.186$. *Nature* **460,** 717–719 (2009).
10. van de Sande, J. *et al.* Stellar kinematics of $z \sim 2$ galaxies and the inside-out growth of quiescent galaxies. *Astrophys. J.* **771,** 85 (2013).
11. Belli, S., Newman, A. B. & Ellis, R. S. Velocity dispersions and dynamical masses for a large sample of quiescent galaxies at $z > 1$: improved measures of the growth in mass and size. *Astrophys. J.* **783,** 117 (2014).
12. Skelton, R. E. *et al.* 3D-HST WFC3-selected photometric catalogs in the five CANDELS/3D-HST fields: photometry, photometric redshifts and stellar masses. Preprint at http://arxiv.org/abs/1403.3689 (2014).
13. van der Wel, A. *et al.* 3D-HST+CANDELS: the evolution of the galaxy size-mass distribution since $z = 3$. *Astrophys. J.* **788,** 28 (2014).
14. Kriek, M. *et al.* An ultra-deep near-infrared spectrum of a compact quiescent galaxy at $z = 2.2$. *Astrophys. J.* **700,** 221–231 (2009).
15. Szomoru, D. *et al.* Confirmation of the compactness of a $z = 1.91$ quiescent galaxy with Hubble Space Telescope's Wide Field Camera 3. *Astrophys. J.* **714,** L244–L248 (2010).
16. Kirkpatrick, A. *et al.* GOODS-Herschel: impact of active galactic nuclei and star formation activity on infrared spectral energy distributions at high redshift. *Astrophys. J.* **759,** 139 (2012).
17. Kennicutt, R. C. Jr. Star formation in galaxies along the Hubble sequence. *Annu. Rev. Astron. Astrophys.* **36,** 189–232 (1998).
18. Chabrier, G. Galactic stellar and substellar initial mass function. *Publ. Astron. Soc. Pacif.* **115,** 763–795 (2003).
19. Wuyts, S. *et al.* Galaxy structure and mode of star formation in the SFR-mass plane from $z \sim 2.5$ to $z \sim 0.1$. *Astrophys. J.* **742,** 96 (2011).
20. da Cunha, E., Charlot, S. & Elbaz, D. A simple model to interpret the ultraviolet, optical and infrared emission from galaxies. *Mon. Not. R. Astron. Soc.* **388,** 1595–1617 (2008).
21. Kennicutt, R. C. Jr. The global Schmidt law in star-forming galaxies. *Astrophys. J.* **498,** 541 (1998).
22. Dekel, A. *et al.* Toy models for galaxy formation versus simulations. *Mon. Not. R. Astron. Soc.* **435,** 999–1019 (2013).
23. Maiolino, R. *et al.* AMAZE. I. The evolution of the mass-metallicity relation at $z > 3$. *Astron. Astrophys.* **488,** 463–479 (2008).
24. Leja, J. *et al.* Exploring the chemical link between local ellipticals and their high-redshift progenitors. *Astrophys. J.* **778,** L24 (2013).
25. Dekel, A. & Burkert, A. Wet disk contraction to galactic blue nuggets and quenching to red nuggets. *Mon. Not. R. Astron. Soc.* **438,** 1870 (2014).
26. Erb, D. K. *et al.* α observations of a large sample of galaxies at $z \sim 2$: implications for star formation in high-redshift galaxies. *Astrophys. J.* **647,** 128–139 (2006).
27. Gilli, R. *et al.* ALMA reveals a warm and compact starburst around a heavily obscured supermassive black hole at $z = 4.75$. *Astron. Astrophys.* **562,** A67 (2014).
28. Wang, W.-H., Barger, A. J. & Cowie, L. L. A Ks and IRAC selection of high-redshift extremely red objects. *Astrophys. J.* **744,** 155 (2012).
29. Barro, G. *et al.* Keck-I MOSFIRE spectroscopy of compact star-forming galaxies at $z \gtrsim 2$: high velocity dispersions in progenitors of compact quiescent galaxies. Preprint at http://arxiv.org/abs/1405.7042 (2014).

Acknowledgements Support from STScI grant GO-1277 is acknowledged.

Author Contributions E.N. obtained the data, led the analysis and the interpretation, and wrote the manuscript. P.v.D. contributed to the analysis and the interpretation. M.F. contributed to the interpretation. I.M. reduced the WFC3 imaging. G.B. and I.M. reduced the grism spectroscopy. K.W. and R.S. led the photometric analysis. All authors commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.N. (erica.nelson@yale.edu).

METHODS

Spectral energy distribution. The candidate forming core was found using the 3D-HST catalogues in the five CANDELS fields. CANDELS is a 902-orbit Hubble Space Telescope program that provides space-based optical and near-infrared imaging across ~900 arcmin^2 (refs 30, 31). Aperture photometry was performed to produce publicly available photometric catalogues and to derive stellar masses[12,14]. Spitzer MIPS 24 μm fluxes were determined using the same methodology as in ref. 32. The derived fluxes are consistent with the public catalogue of ref. 33. Using the 24 μm data as position priors, we measure the 100–500 μm fluxes from the ultra-deep Herschel imaging in GOODS-North[34]. In sum, the rest-frame ultraviolet–optical data come from HST/ACS, HST/WFC3 and ground-based optical telescopes; the rest-frame near-infrared data are from Spitzer/IRAC; the mid-infrared point is from Spitzer/MIPS; and the far-infrared data are from Herschel/PACS and SPIRE.

Keck spectroscopy. We observed GOODS-N-774 with the near infrared spectrograph (NIRSPEC) on the Keck I telescope in the K band, on 11 January 2014. The total integration time was 6,000 s. We used the low-dispersion mode with a slit width of 0.7″, giving a spectral resolution of $\sigma_{\rm instr} = 6.1$ Å in the rest frame. We fitted a Gaussian to the Hα $\lambda = 6{,}563$ Å and [N II] $\lambda = 6{,}548$ and 6,584 Å emission lines simultaneously and corrected for the instrumental resolution. The uncertainty in the derived properties was determined by refitting the model with empirical realizations of the noise.

HST grism spectroscopy. A WFC3/G141 grism spectrum of the object was obtained as part of the 3D-HST survey[35]. 3D-HST is a near-infrared slitless spectroscopic Treasury programme. We examined the grism spectrum after measuring a secure redshift from the Keck/NIRSPEC spectrum. The redshifted [O II], Hβ and [O III] lines are detected with a significance of 1.5σ–2.5σ. GOODS-N-774 has optical emission line ratios [O III]/H$\beta = 1.2 \pm 0.9$ and [N II]/H$\alpha = 0.4 \pm 0.1$, suggesting a level of gas excitation that is slightly higher than the locus of star-forming galaxies in the local Universe[36] but at the low end for star-forming galaxies at $z \approx 2$ (ref. 37) in the diagnostic BPT diagram.

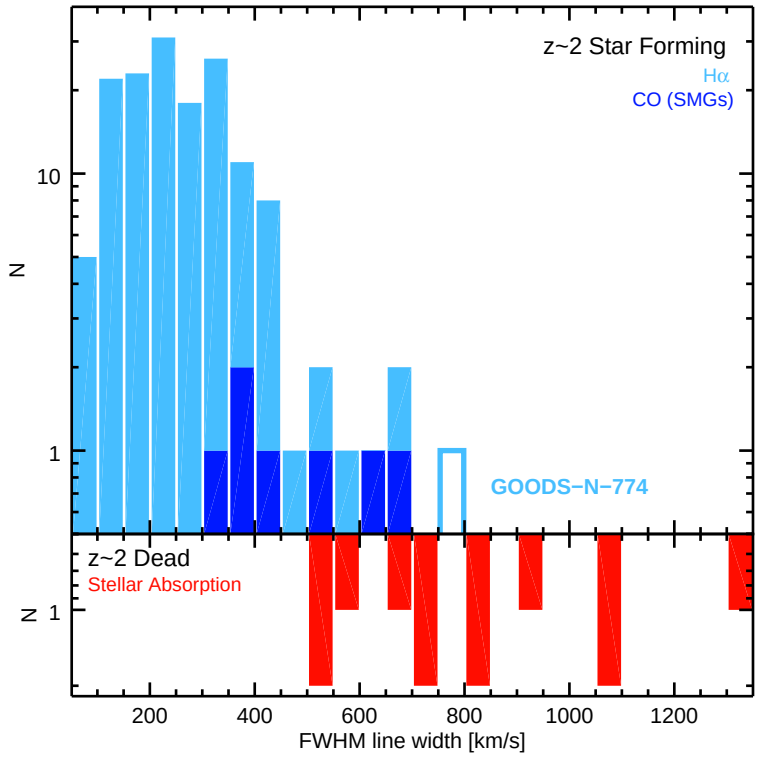
X-ray constraints. GOODS-N-774 is in the Chandra Deep Field North, which has been observed for a total of ~2 Ms with the Chandra X-ray satellite. The exposure time at the location of GOODS-N-774 is 1.22 Ms. The galaxy is not in the publicly available point-source catalogue of this field[38]. There are seven counts in a 3″ aperture centred on the object location in the full-band (0.5–8 keV) X-ray image, fully consistent with the counts in random apertures in regions with the same exposure time. Using the s.d. of the counts in random apertures, we derive a 2 s.d. upper limit of six counts for the X-ray flux of GOODS-N-774. Using PIMMS v4.6b, we derive a rest-frame 2–10 keV flux limit of $F_{\rm X} < 2.9 \times 10^{-17}$ erg s^{-1} cm^{-2}, corresponding to a luminosity of $L_{\rm X} < 1.2 \times 10^{42}$ erg s^{-1}. We conclude that there is no evidence for an AGN in GOODS-N-774. The upper limit is consistent with the star formation rate of the galaxy[39].

Gas column density. We derive the gas surface density using the Kennicutt–Schmidt law[21]:

$$\Sigma_{\rm SFR} = (2.5 \pm 0.7) \times 10^{-4} \left(\frac{\Sigma_{\rm gas}}{1\,M_{\odot}\ {\rm pc}^{-2}} \right)^{1.4 \pm 0.15} M_{\odot}\ {\rm yr}^{-1}\ {\rm kpc}^{-2}$$

Dynamical mass. We define dynamical mass as $M_{\rm dyn} = k(n)\sigma^2 r_{\rm e}/G$, with the constant $k(n)$ depending on the Sérsic index: $k(n) = 8.87 - 0.831n + 0.0241n^2$ (ref. 40). GOODS-N-774 has a Sérsic index of $n = 2.9$; the comparison samples of compact quiescent galaxies at $z \approx 2$ (refs 1, 41–45) and SDSS galaxies with $0.058 < z < 0.060$ have median indices of $n = 3.2$ and $n = 4.1$, respectively.

30. Grogin, N. *et al.* CANDELS: the Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey. *Astrophys. J. Suppl. Ser.* **197,** 35 (2011).
31. Koekemoer, A. *et al.* CANDELS: the Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey — the Hubble Space Telescope observations, imaging data products and mosaics. *Astrophys. J. Suppl. Ser.* **197,** 36 (2011).
32. Whitaker, K. *et al.* The NEWFIRM medium-band survey: photometric catalogs, redshifts, and the bimodal color distribution of galaxies out to $z \sim 3$. *Astrophys. J.* **735,** 86 (2011).
33. Lutz, D. *et al.* PACS Evolutionary Probe (PEP) - a Herschel key program. *Astron. Astrophys.* **532,** A90 (2011).
34. Elbaz, D. *et al.* GOODS-Herschel: an infrared main sequence for star-forming galaxies. *Astron. Astrophys.* **533,** A119 (2011).
35. Brammer, G. B. *et al.* 3D-HST: a wide-field grism spectroscopic survey with the Hubble Space Telescope. *Astrophys. J.* **200,** 13 (2012).
36. Tremonti, C. *et al.* The origin of the mass-metallicity relation: insights from 53,000 star-forming galaxies in the Sloan Digital Sky Survey. *Astrophys. J.* **613,** 898 (2004).
37. Steidel, C. *et al.* Strong nebular line ratios in the spectra of $z \sim 2$–3 star-forming galaxies: first results from KBSS-MOSFIRE. *Astrophys. J.* (submitted); preprint at http://arxiv.org/abs/1405.5473 (2014).
38. Alexander, D. *et al.* The Chandra Deep Field North Survey. XIII. 2 Ms point-source catalogs. *Astron. J.* **126,** 539 (2003).
39. Grimm, H.-J., Gilfanov, M. & Sunyaev, R. High-mass X-ray binaries as a star formation rate indicator in distant galaxies. *Mon. Not. R. Astron. Soc.* **339,** 793 (2003).
40. Bertin, G., Ciotti, L. & Del Principe, M. Weak homology of elliptical galaxies. *Astron. Astrophys.* **386,** 149–168 (2002).
41. Trujillo, I. *et al.* The size evolution of galaxies since $z \sim 3$: combining SDSS, GEMS, and FIRES. *Astrophys. J.* **650,** 18–41 (2006).
42. Toft, S. *et al.* Hubble Space Telescope and Spitzer imaging of red and blue galaxies at $z \sim 2.5$: a correlation between size and star formation activity from compact quiescent galaxies to extended star-forming galaxies. *Astrophys. J.* **671,** 285–302 (2007).
43. van Dokkum, P. G. *et al.* Confirmation of the remarkable compactness of massive quiescent galaxies at $z \sim 2.3$: early-type galaxies did not form in a simple monolithic collapse. *Astrophys. J.* **677,** L5–L8 (2008).
44. Cimatti, A. *et al.* GMASS ultradeep spectroscopy of galaxies at $z \sim 2$. II. Superdense passive galaxies: how did they form and evolve? *Astron. Astrophys.* **482,** 21–42 (2008).
45. Newman, A. B., Ellis, R. S., Treu, T. & Bundy, K. Keck spectroscopy of $z > 1$ field spheroidals: dynamical constraints on the growth rate of red ''nuggets''. *Astrophys. J.* **717,** L103–L107 (2010).
46. Förster Schreiber, N. M. *et al.* The SINS survey: SINFONI integral field spectroscopy of $z \sim 2$ star-forming galaxies. *Astrophys. J.* **706,** 1364–1428 (2009).



Extended Data Figure 1 | Linewidths of $z \approx 2$ star-forming and quiescent galaxies. The linewidth of GOODS-N-774 (open box) is among the highest measured for a normal star-forming galaxy at high redshift in Hα emission[26,46] (light blue) or CO emission[4] (SMGs; dark blue). The gas velocity dispersion is similar to the median stellar velocity dispersion of 304 km s^{-1} in a sample of quiescent galaxies at $z = 1.5$–2.2 with median mass of $1.9 \times 10^{11} M_{\odot}$ (refs 8–11; red).

LETTER

doi:10.1038/nature13762

A supermassive black hole in an ultra-compact dwarf galaxy

Anil C. Seth[1], Remco van den Bosch[2], Steffen Mieske[3], Holger Baumgardt[4], Mark den Brok[1], Jay Strader[5], Nadine Neumayer[2,6], Igor Chilingarian[7,8], Michael Hilker[6], Richard McDermid[9,10], Lee Spitler[9,10], Jean Brodie[11], Matthias J. Frank[12] & Jonelle L. Walsh[13]

Ultra-compact dwarf galaxies are among the densest stellar systems in the Universe. These systems have masses of up to 2×10^8 solar masses, but half-light radii of just 3–50 parsecs[1]. Dynamical mass estimates show that many such dwarfs are more massive than expected from their luminosity[2]. It remains unclear whether these high dynamical mass estimates arise because of the presence of supermassive black holes or result from a non-standard stellar initial mass function that causes the average stellar mass to be higher than expected[3,4]. Here we report adaptive optics kinematic data of the ultra-compact dwarf galaxy M60-UCD1 that show a central velocity dispersion peak exceeding 100 kilometres per second and modest rotation. Dynamical modelling of these data reveals the presence of a supermassive black hole with a mass of 2.1×10^7 solar masses. This is 15 per cent of the object's total mass. The high black hole mass and mass fraction suggest that M60-UCD1 is the stripped nucleus of a galaxy. Our analysis also shows that M60-UCD1's stellar mass is consistent with its luminosity, implying a large population of previously unrecognized supermassive black holes in other ultra-compact dwarf galaxies[2].

The object M60-UCD1 is the brightest ultracompact dwarf galaxy (UCD) currently known[5], with a V-band luminosity $L_V = 4.1 \times 10^7\ L_\odot$ (where $L_\odot$ is the solar luminosity) and effective radius $r_e = 24$ pc. It lies at a projected distance of 6.6 kpc from the centre of the massive elliptical galaxy M60 (Fig. 1), and 16.5 Mpc from us[6]. We obtained integral field spectroscopic data between 2 μm and 2.4 μm of M60-UCD1 with the near-infrared integral field spectrograph on the Gemini North telescope. The high-spatial-resolution data obtained using laser guide-star adaptive optics provides a clear detection of the supermassive black hole. Modelling of the deep carbon dioxide (CO) absorption bandheads at 2.3 μm enables us to measure the motions of stars at many different points across M60-UCD1. These kinematic measurements are shown in Fig. 2. Two features are particularly notable: (1) the dispersion is strongly peaked, with the central dispersion rising above 100 km s^{-1} and dropping outwards to ~50 km s^{-1}, (2) rotation is clearly seen, with a peak amplitude of 40 km s^{-1}.

The stellar kinematics can be used to constrain the distribution of mass within M60-UCD1. This includes being able to test whether the mass traces light, or whether a supermassive black hole is required to explain the central velocity dispersion peak. We combined the stellar kinematics and imaging from the Hubble Space Telescope with self-consistent Schwarzschild models[7–9] to constrain the black hole mass and the mass-to-light ratio (M/L, in solar units), shown in Fig. 3. We measure a black-hole mass of $2.1^{+1.4}_{-0.7} \times 10^7\ M_\odot$ and a g-band $M/L = 3.6 \pm 1$ with errors giving 1σ confidence intervals (2σ and 3σ contours are shown in Fig. 3). The total stellar mass is $1.2 \pm 0.4 \times 10^8\ M_\odot$, where $M_\odot$ is the solar mass. The best-fitting constant-M/L model with no black hole is ruled out with >99.99% confidence.

M60-UCD1 is the lowest-mass system known to host a supermassive black hole ($> 10^6\ M_\odot$), including systems with dynamical black hole estimates or with broad-line active galactic nuclei[10,11]. There have been tentative detections of approximately $10^4\ M_\odot$ black holes in lower-mass clusters[12,13]. These detections remain controversial[14,15] and the intermediate-mass black holes, if present, form a much smaller fraction of the total cluster mass than found in M60-UCD1. Of the 75 galaxies with reliable dynamical black hole mass measurements, only one other galaxy has a black hole mass fraction as high as that of M60-UCD1[10,16]. A luminous and variable X-ray source was previously detected in M60-UCD1 with a maximum luminosity of $L_X = 1.3 \times 10^{38}$ ergs s^{-1} (ref. 5). This luminosity suggests the black hole is accreting material at a rate typical of black holes in larger, more-massive early-type galaxies as well as other nearby galaxies with absorption-line-dominated optical spectra[17,18].

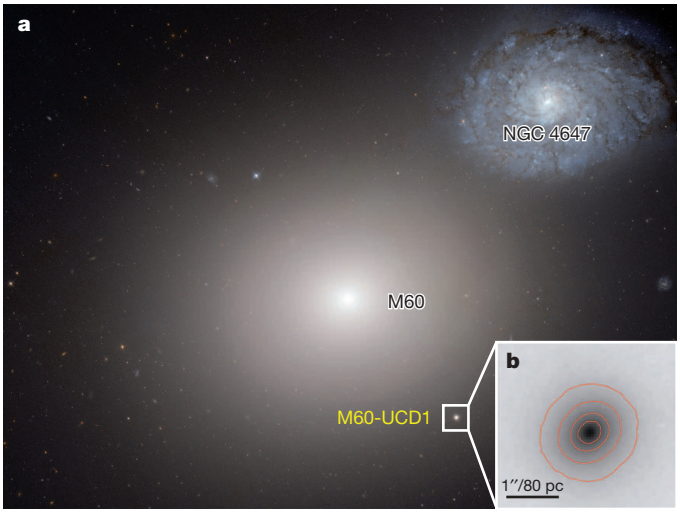


Figure 1 | Hubble Space Telescope image of the M60–NGC 4647 system. M60-UCD1 is the nearly point-like image in the bottom right of **a** (boxed). The discovery of a black hole in M60-UCD1 provides evidence that it is the tidally stripped nucleus of a once larger galaxy. We note that NGC 4647 is at approximately the same distance from Earth as M60 but the two galaxies are not yet strongly interacting. **b**, A zoomed version of the g-band image of M60-UCD1 with contours showing the surface brightness in intervals of one magnitude per square arcsecond. The image is from NASA/ESA.

[1]Department of Physics and Astronomy, University of Utah, 115 South 1400 East, Salt Lake City, Utah 84112, USA. [2]Max-Planck Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany. [3]European Southern Observatory, Alonso de Cordova 3107, Vitacura, Santiago, 7630355, Chile. [4]School of Mathematics and Physics, University of Queensland, St Lucia, Queensland 4072, Australia. [5]Department of Physics and Astronomy, Michigan State University, East Lansing, Michigan 48824, USA. [6]European Southern Observatory, Karl-Schwarzschild-Strasse 2, 85748 Garching bei München, Germany. [7]Smithsonian Astrophysical Observatory, 60 Garden Street MS09, Cambridge, Massachusetts 02138, USA. [8]Sternberg Astronomical Institute, Moscow State University, 13 Universitetski prospect, Moscow 119992, Russia. [9]Australian Astronomical Observatory, 105 Delhi Road, Sydney, New South Wales 2113, Australia. [10]Department of Physics and Astronomy, Macquarie University, Sydney, New South Wales 2109, Australia. [11]University of California Observatories and Department of Astronomy and Astrophysics, University of California, Santa Cruz, California 95064, USA. [12]Landessternwarte, Zentrum für Astronomie der Universität Heidelberg, Königsstuhl 12, D-69117 Heidelberg, Germany. [13]Department of Astronomy, The University of Texas at Austin, 1 University Station C1400, Austin, Texas 78712, USA.

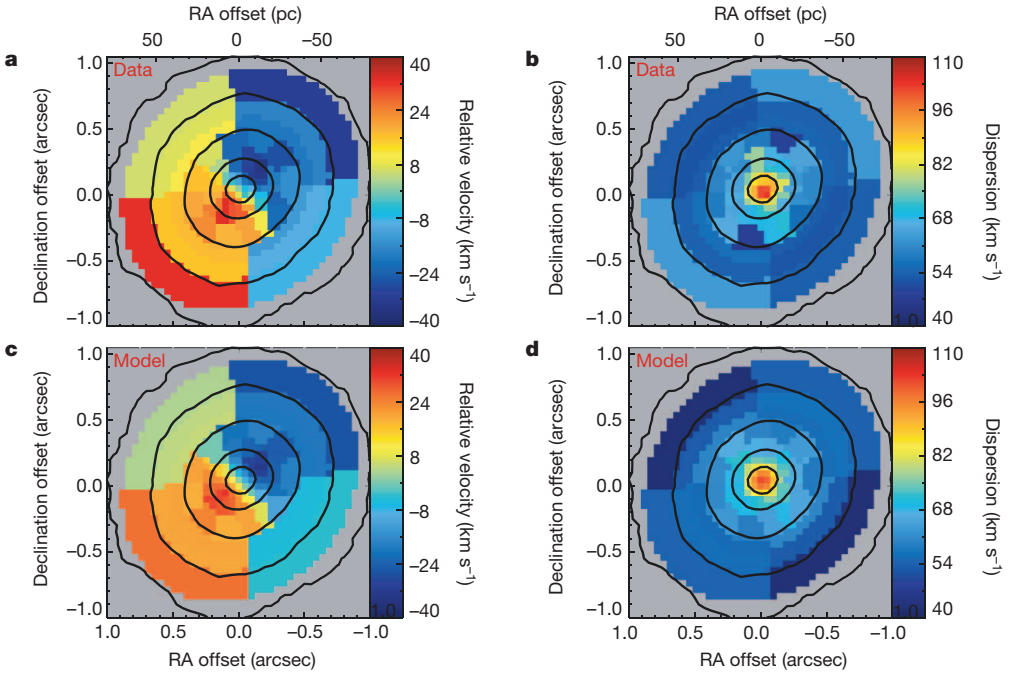


Figure 2 | Stellar kinematic maps of M60-UCD1 showing clear rotation and a dispersion peak. **a** and **b** show the measured radial velocities (bulk motions towards and away from us) and velocity dispersions (random motions) of the stars in M60-UCD1 with typical errors of 6 km s^{-1}. Black contours show isophotes in the K-band stellar continuum. Kinematics are determined in each individual pixel near the centre, but at larger radii the data were binned to increase the signal-to-noise ratio and enable kinematic measurements. **c** and **d** show the best-fitting dynamical model; a black hole is required to replicate the central dispersion peak.

UCDs are thought to be either the most-massive globular star clusters[19] or nuclei of larger galaxies that have been tidally stripped[20,21]. The supermassive black hole that we have found at the centre of M60-UCD1 provides strong evidence that it is a stripped nucleus of a once larger galaxy. While it is possible that dense star clusters can form black holes, these are expected to contain only a small fraction of the cluster's mass[22]. Star clusters at the centre of galaxy nuclei, on the other hand, are known to host black holes with very high mass fractions[23]. Thus M60-UCD1 is the first individual UCD with explicit evidence for being a tidally stripped nucleus.

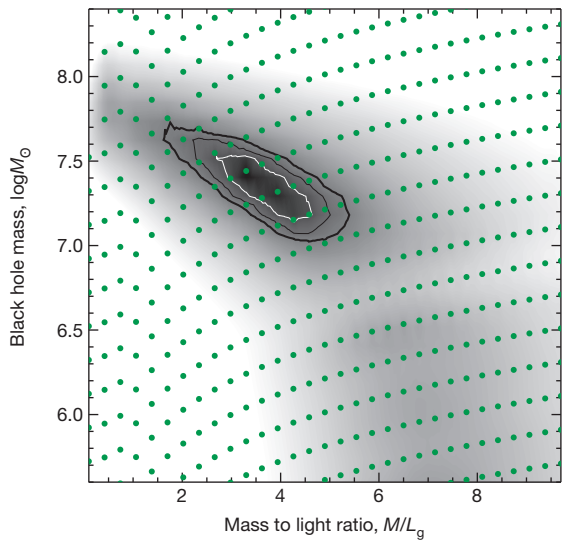


Figure 3 | Dynamical modelling results show the presence of a supermassive black hole. The figure shows goodness-of-fit contours for the dynamical models of M60-UCD1 with two parameters, the g-band mass-to-light ratio and the black hole mass. The contours indicate 1σ (white), 2σ, 3σ (black, thick) confidence levels for two degrees of freedom. Green dots indicate discrete values of the mass-to-light ratio and the black hole mass at which models were fitted to the data.

We can estimate the properties of M60-UCD1's progenitor galaxy, assuming that they follow scaling relations of present-day unstripped galaxies. Using the known scaling between black hole mass and bulge mass[10], we find a host bulge mass of $7^{+4}_{-3} \times 10^9\ M_\odot$. Bulge masses are also known to correlate with the masses of their nuclear star clusters[24]. In M60-UCD1, the surface brightness profile has two clear components[5], and we identify the central component as the progenitor nuclear star cluster[21] with mass $6.1 \pm 1.6 \times 10^7\ M_\odot$. This translates to a predicted bulge mass of $1.8 \pm 0.4 \times 10^{10}\ M_\odot$. Thus, two independent scaling relations suggest that the progenitor bulge mass is about $10^{10}\ M_\odot$. Given M60-UCD1's cluster environment, its progenitor was probably a lower-mass elliptical galaxy that was then stripped by the massive elliptical galaxy M60, which lies at a current projected distance of just 6.6 kpc. We have run simulations that show that it is feasible to produce M60-UCD1 by stripping an elliptical galaxy progenitor of approximately $10^{10}\ M_\odot$ on a fairly radial orbit (see Supplementary Information and Extended Data Fig. 6). We note that current elliptical galaxies of approximately $10^{10}\ M_\odot$ have nuclear star cluster sizes, luminosities and colours consistent with the inner component of M60-UCD1[25,26].

The detection of a supermassive black hole in M60-UCD1 may be just the tip of the iceberg of the UCD black hole population. Measurements of the integrated velocity dispersion in almost all UCDs with masses above $10^7\ M_\odot$ yield dynamical mass estimates that are too high to be accounted for by a normal stellar population without a massive central black hole[2]. Unlike these previous dynamical mass estimates, our dynamical modelling can separate out the gravitational influence of the black hole from the contribution of the stars. The models show that M60-UCD1's stellar populations appear normal. The mass-to-light ratio of a stellar population characterizes the average mass of its stars. In M60-UCD1 we measured a stellar dynamical mass-to-light ratio in the g band of $M/L_g = 3.6 \pm 1.0$ in solar units (1σ errors). This mass-to-light ratio is consistent with the stellar populations seen in lower-mass globular clusters[27] and models with a normal (Milky Way) initial mass function[28]. It is also lower than the integrated dynamical mass-to-light ratio estimates

in 18 of 19 UCDs above $10^7\ M_\odot$ (ref. 2). The low stellar mass-to-light ratio in M60-UCD1 is inconsistent with proposed scenarios in which a density-dependent initial mass function yields higher than normal mass-to-light ratios[4]. Without such a mechanism, the simplest explanation for the high dynamical mass estimates in massive UCDs is that most host supermassive black holes just as M60-UCD1 does.

There is also more limited evidence for enhanced dynamical mass-to-light ratios in lower-mass UCDs. About half of UCDs with masses between $3 \times 10^6\ M_\odot$ and $10^7\ M_\odot$ have higher inferred mass-to-light ratios than we see in M60-UCD1[2]. Taken in combination with tentative black hole detections of about $10^4\ M_\odot$ in Local Group globular clusters[12,13] (which do not have increased M/L values owing to much lower black hole mass fractions), these observations suggest that some lower-mass UCDs may also host relatively massive black holes.

Finally, we estimate what the total population of UCD black holes might be in the local Universe. The most complete sample of known UCDs is in the Fornax cluster. Comparing the UCD population to the population of galaxies in Fornax likely to host massive black holes, we find that UCDs may more than double the number of black holes (see Supplementary Information). Thus UCD black holes could represent a large increase in the massive black hole number density in the local Universe. Future work can test this hypothesis. We have ongoing observing programmes to obtain similar observations to the ones presented here in four additional massive UCDs, and in the most massive star clusters in the Local Group. However, dynamical detection of black holes will be challenging in all but the brightest and nearest of objects, so accretion signatures[29] or tidal disruption events[30] may represent the best possibility for detecting black holes in less-massive UCDs.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 June; accepted 29 July 2014.

1. Brodie, J. P., Romanowsky, A. J., Strader, J. & Forbes, D. A. The relationships among compact stellar systems: a fresh view of ultracompact dwarfs. *Astron. J.* **142,** 199 (2011).
2. Mieske, S. *et al.* On central black holes in ultra-compact dwarf galaxies. *Astron. Astrophys.* **558,** A14 (2013).
3. Frank, M. J. *et al.* Spatially resolved kinematics of an ultracompact dwarf galaxy. *Mon. Not. R. Astron. Soc.* **414,** L70–L74 (2011).
4. Dabringhausen, J., Kroupa, P., Pflamm-Altenburg, J. & Mieske, S. Low-mass X-ray binaries indicate a top-heavy stellar initial mass function in ultracompact dwarf galaxies. *Astrophys. J.* **747,** 72 (2012).
5. Strader, J. *et al.* The densest galaxy. *Astrophys. J.* **775,** L6 (2013).
6. Blakeslee, J. P. *et al.* The ACS Fornax cluster survey. V. Measurement and recalibration of surface brightness fluctuations and a precise value of the Fornax–Virgo relative distance. *Astrophys. J.* **694,** 556–572 (2009).
7. Schwarzschild, M. A numerical model for a triaxial stellar system in dynamical equilibrium. *Astrophys. J.* **232,** 236–247 (1979).
8. van den Bosch, R. C. E., van de Ven, G., Verolme, E. K., Cappellari, M. & de Zeeuw, P. T. Triaxial orbit based galaxy models with an application to the (apparent) decoupled core galaxy NGC 4365. *Mon. Not. R. Astron. Soc.* **385,** 647–666 (2008).
9. van den Bosch, R. C. E. & de Zeeuw, P. T. Estimating black hole masses in triaxial galaxies. *Mon. Not. R. Astron. Soc.* **401,** 1770–1780 (2010).
10. Kormendy, J. & Ho, L. C. Coevolution (or not) of supermassive black holes and host galaxies. *Annu. Rev. Astron. Astrophys.* **51,** 511–653 (2013).
11. Reines, A. E., Greene, J. E. & Geha, M. Dwarf galaxies with optical signatures of active massive black holes. *Astrophys. J.* **775,** 116 (2013).
12. Gebhardt, K., Rich, R. M. & Ho, L. C. An intermediate-mass black hole in the globular cluster G1: improved significance from new Keck and Hubble Space Telescope observations. *Astrophys. J.* **634,** 1093–1102 (2005).
13. Jalali, B. *et al.* A dynamical N-body model for the central region of ω Centauri. *Astron. Astrophys.* **538,** A19 (2012).
14. van der Marel, R. P. & Anderson, J. New limits on an intermediate-mass black hole in Omega Centauri. II. Dynamical models. *Astrophys. J.* **710,** 1063–1088 (2010).
15. Miller-Jones, J. C. A. *et al.* The absence of radio emission from the globular cluster G1. *Astrophys. J.* **755,** L1 (2012).
16. van den Bosch, R. C. E. *et al.* An over-massive black hole in the compact lenticular galaxy NGC1277. *Nature* **491,** 729–731 (2012).
17. Gallo, E. *et al.* AMUSE-Virgo. II. Down-sizing in black hole accretion. *Astrophys. J.* **714,** 25–36 (2010).
18. Ho, L. C. Nuclear activity in nearby galaxies. *Annu. Rev. Astron. Astrophys.* **46,** 475–539 (2008).
19. Mieske, S., Hilker, M. & Misgeld, I. The specific frequencies of ultra-compact dwarf galaxies. *Astron. Astrophys.* **537,** A3 (2012).
20. Drinkwater, M. J. *et al.* A class of compact dwarf galaxies from disruptive processes in galaxy clusters. *Nature* **423,** 519–521 (2003).
21. Pfeffer, J. & Baumgardt, H. Ultra-compact dwarf galaxy formation by tidal stripping of nucleated dwarf galaxies. *Mon. Not. R. Astron. Soc.* **433,** 1997–2005 (2013).
22. Portegies Zwart, S. F., Baumgardt, H., Hut, P., Makino, J. & McMillan, S. L. W. Formation of massive black holes through runaway collisions in dense young star clusters. *Nature* **428,** 724–726 (2004).
23. Graham, A. W. & Spitler, L. R. Quantifying the coexistence of massive black holes and dense nuclear star clusters. *Mon. Not. R. Astron. Soc.* **397,** 2148–2162 (2009).
24. Ferrarese, L. *et al.* A fundamental relation between compact stellar nuclei, supermassive black holes, and their host galaxies. *Astrophys. J.* **644,** L21–L24 (2006).
25. Côté, P. *et al.* The ACS Virgo cluster survey. VIII. The nuclei of early-type galaxies. *Astrophys. J.* **165** (suppl.), 57–94 (2006).
26. Turner, M. L. *et al.* The ACS Fornax cluster survey. VI. The nuclei of early-type galaxies in the Fornax cluster. *Astrophys. J.* **203** (suppl.), 5 (2012).
27. Strader, J., Caldwell, N. & Seth, A. C. Star clusters in M31. V. Internal dynamical trends: some troublesome, some reassuring. *Astron. J.* **142,** 8 (2011).
28. Bastian, N., Covey, K. R. & Meyer, M. R. A universal stellar initial mass function? A critical look at variations. *Annu. Rev. Astron. Astrophys.* **48,** 339–389 (2010).
29. Gültekin, K., Cackett, E. M., King, A. L., Miller, J. M. & Pinkney, J. Low-mass AGNs and their relation to the fundamental plane of black hole accretion. *Astrophys. J.* **788,** L22 (2014).
30. Miller, M. C., Farrell, S. A. & Maccarone, T. J. A wind accretion model for HLX-1. *Astrophys. J.* **788,** 116 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was based on observations obtained at the Gemini Observatory, which is operated by the Association of Universities for Research in Astronomy under a cooperative agreement with the NSF on behalf of the Gemini partnership: the National Science Foundation (United States), the National Research Council (Canada), CONICYT (Chile), the Australian Research Council (Australia), Ministério da Ciência, Tecnologia e Inovação (Brazil) and Ministerio de Ciencia, Tecnología e Innovación Productiva (Argentina). Work on this paper by A.C.S. was supported by NSF CAREER grant AST-1350389. J.L.W. is supported by an NSF Astronomy and Astrophysics Postdoctoral Fellowship under award number 1102845. J.B. is supported by NSF grant AST-1109878. M.J.F. is supported by German Research Foundation grant Ko4161/1. I.C. acknowledges support from the Russian Science Foundation grant 14-22-00041.

Author Contributions All authors helped with interpretation of the data and provided comments on the manuscript. A.C.S. planned observations, reduced and analysed the data and was the primary author of the text. R.v.d.B. created dynamical models and contributed text. S.M. contributed text. H.B. ran tidal stripping simulations. M.d.B. created dynamical models and analysed model results. J.S., N.N., and R.M. helped to plan the observations. I.C. helped verify kinematic measurements. M.H. and L.S. helped with the compilation of UCD numbers.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.C.S. (aseth@astro.utah.edu).

LETTER

doi:10.1038/nature13705

Aridification of the Sahara desert caused by Tethys Sea shrinkage during the Late Miocene

Zhongshi Zhang[1,2], Gilles Ramstein[3], Mathieu Schuster[4], Camille Li[5], Camille Contoux[1] & Qing Yan[2]

It is widely believed that the Sahara desert is no more than ~2–3 million years (Myr) old[1], with geological evidence showing a remarkable aridification of north Africa at the onset of the Quaternary ice ages[2–4]. Before that time, north African aridity was mainly controlled by the African summer monsoon (ASM)[5–8], which oscillated with Earth's orbital precession cycles. Afterwards, the Northern Hemisphere glaciation added an ice volume forcing on the ASM, which additionally oscillated with glacial–interglacial cycles[2]. These findings led to the idea that the Sahara desert came into existence when the Northern Hemisphere glaciated ~2–3 Myr ago. The later discovery, however, of aeolian dune deposits ~7 Myr old[9] suggested a much older age, although this interpretation is hotly challenged[1] and there is no clear mechanism for aridification around this time. Here we use climate model simulations to identify the Tortonian stage (~7–11 Myr ago) of the Late Miocene epoch as the pivotal period for triggering north African aridity and creating the Sahara desert. Through a set of experiments with the Norwegian Earth System Model[10] and the Community Atmosphere Model[11], we demonstrate that the African summer monsoon was drastically weakened by the Tethys Sea shrinkage during the Tortonian, allowing arid, desert conditions to expand across north Africa. Not only did the Tethys shrinkage alter the mean climate of the region, it also enhanced the sensitivity of the African monsoon to orbital forcing, which subsequently became the major driver of Sahara extent fluctuations. These important climatic changes probably caused the shifts in Asian and African flora and fauna observed during the same period[4,12–14], with possible links to the emergence of early hominins in north Africa[15,16].

The Sahara desert, the largest non-polar desert in the world today (~9,400,000 km^2), is widely believed to have appeared during the past 2 or 3 Myr (ref. 1). A wealth of terrestrial and marine evidence indicates a remarkable aridification across north Africa since the onset of the Quaternary ice ages[2–4]. The arid conditions have not been constant. North African aridity varies with the strength of the ASM, which is influenced by the ~20-kyr precession cycle[2,5–8] as well as by global ice volume fluctuations that follow ~40-kyr or ~100-kyr cycles[2,6]. When orbital precession maximizes Northern Hemisphere summer insolation, or when the ice volume is small, a strong ASM brings more moisture into north Africa from the tropical Atlantic[5,6]. The increased precipitation limits the extent of the desert, even replacing it at times with a 'green Sahara' landscape of vegetation, rivers and lakes that favours occupation by fauna and humans. The relationship between north African aridity, the ASM and glacial–interglacial cycles has been used as further support for the notion that the Sahara desert first appeared at the same time as the ice ages started (~2–3 Myr ago).

Recently, geological evidence for an earlier appearance of the Sahara desert has been mounting. Dust flux[2,3] and terrestrial vegetation archives[4] from marine sediments indicate the existence of arid periods in north Africa since ~8 Myr ago. On land, aeolian dune deposits in the northern Chad basin suggest that desert conditions existed as early as 7 Myr ago[9]. In the same area, several large palaeolake recurrences recorded between 3 and 7 Myr ago also reveal arid–humid cycles[17]. This evidence for the early onset of Saharan aridity is, however, hotly challenged[1], and although the appearance of the Sahara desert 2–3 Myr ago can be tied to the onset of the Quaternary ice ages, there are no candidate explanations for its earlier appearance.

To investigate possible explanations for the onset of Saharan aridity, we simulate climate change in north Africa on geological timescales over the past 30 Myr. First, we simulate the climate of the Late Oligocene, the Early Miocene, the Late Miocene and today, using the coupled low-resolution version of the Norwegian Earth System Model (NorESM-L)[10]. Then, with a high-resolution version of the atmospheric component of NorESM-L, the Community Atmosphere Model version 4 (CAM4)[11], we perform sensitivity experiments to identify the dominant drivers for the simulated climate changes in north Africa. (See Methods and Extended Data for a detailed description of the boundary conditions, experimental design and a discussion of possible model bias and uncertainties.)

In the coupled NorESM-L simulations, north Africa experiences a pronounced aridification from the Early Miocene to the Late Miocene (Fig. 1). In the Late Oligocene and the Early Miocene experiments, north Africa is dominated by a semiarid steppe climate (according to the Köppen climate classification) with only restricted areas of arid desert climate (Fig. 1a, b). In the Late Miocene experiment, the arid desert climate expands across much of north Africa (Fig. 1c), with a greater resemblance to today's conditions (Fig. 1d). The aridification is due to a reduction in north African precipitation (Fig. 2) from annual values (regionally averaged) ~400 mm in the Late Oligocene and the Early Miocene experiments to <200 mm (a criterion for classifying arid climates) in the Late Miocene experiments (Fig. 2e, f).

The simulated aridification is not caused by changes in vegetation, orbital parameters or atmospheric concentrations of greenhouse gas. Vegetation and orbital changes can influence precipitation in north Africa[18], but these are identical in all the coupled palaeoclimate experiments and hence cannot be responsible for the simulated aridification. Another possible cause is the 'greenhouse effect' associated with atmospheric CO_2, whose concentrations decrease from the Oligocene to the Miocene. However, when the experiments are repeated with the atmospheric CO_2 concentrations held constant, north Africa experiences reduced precipitation (Fig. 2c, d) and aridification (Fig. 2f) in the Late Miocene experiment.

The remaining possibility for explaining the simulated aridification is tectonic changes, including changes in orography, bathymetry and, most importantly, land–sea distribution. Orographic changes, mainly mountain uplift, can be excluded. Previous studies have shown that the major orogeneses occurring during the Neogene (the past 23 Myr)—that is, the uplift of eastern and southern Africa[19] and the Tibetan Plateau[20,21]—do not have a significant impact on north African climate[22–24]. Changes in bathymetry that influence regional climate are often linked to changes in land–sea distribution. Between the Early Miocene and the Late Miocene,

[1]Bjerknes Centre for Climate Research, Uni Research Climate, Allégaten 70, 5007 Bergen, Norway. [2]Nansen-Zhu International Research Centre, Institute of Atmospheric Physics, Chinese Academy of Sciences, 100029 Beijing, China. [3]Laboratoire des Sciences du Climat et de l'Environnement/IPSL, CEA-CNRS-UVSQ, UMR8212, Orme des Merisiers, CE Saclay, 91191 Gif-sur-Yvette Cedex, France. [4]Institut de Physique du Globe de Strasbourg (UMR 7516), École et Observatoire des Sciences de la Terre, CNRS and Université de Strasbourg, 1 rue Blessig, 67084 Strasbourg Cedex, France. [5]Bjerknes Centre for Climate Research, Geophysical Institute, University of Bergen, Allégaten 70, 5007 Bergen, Norway.

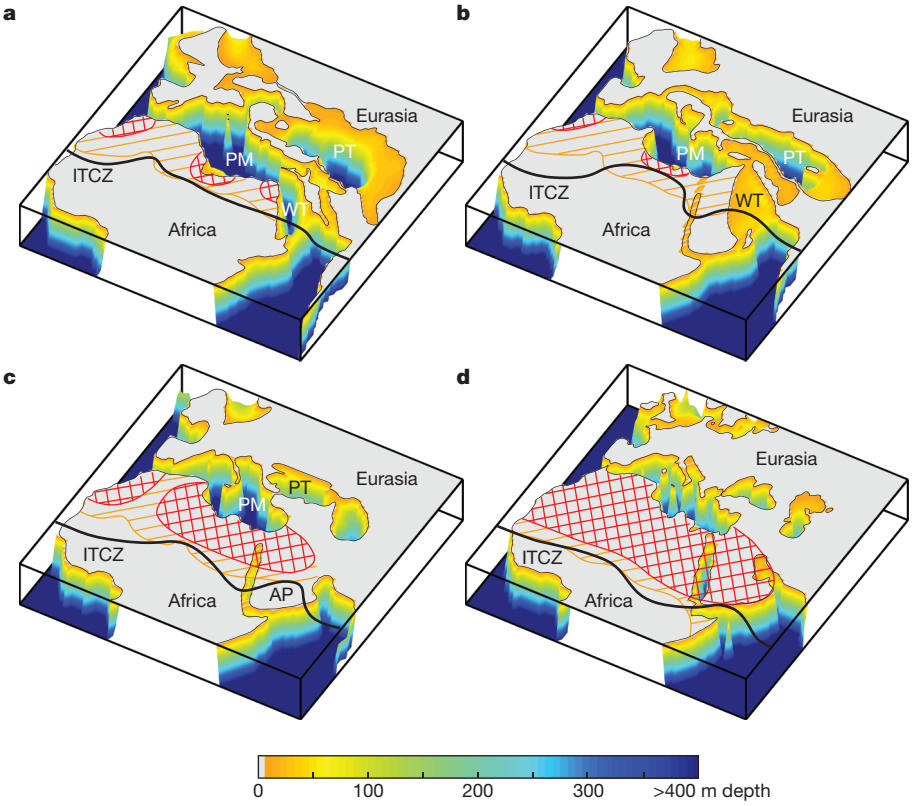


Figure 1 | Tethys palaeogeography and schematic north African palaeoclimate. Late Oligocene (**a**), Early Miocene (**b**), Late Miocene (**c**) and modern (**d**) time slices. The reconstructed Tethys bathymetry is based on the palaeogeography map created in refs 26 and 27, showing the West Tethys Sea (WT), Paratethys Sea (PT), proto-Mediterranean Sea (PM) and Arabian peninsula (AP). Hatched areas show semiarid steppe climate (orange) and arid desert climate (red) from simulations, according to the Köppen climate classification. The black lines show the simulated climatological mean Intertropical Convergence Zone (ITCZ) in summer (June to August), which is a measure of the intensity of the ASM.

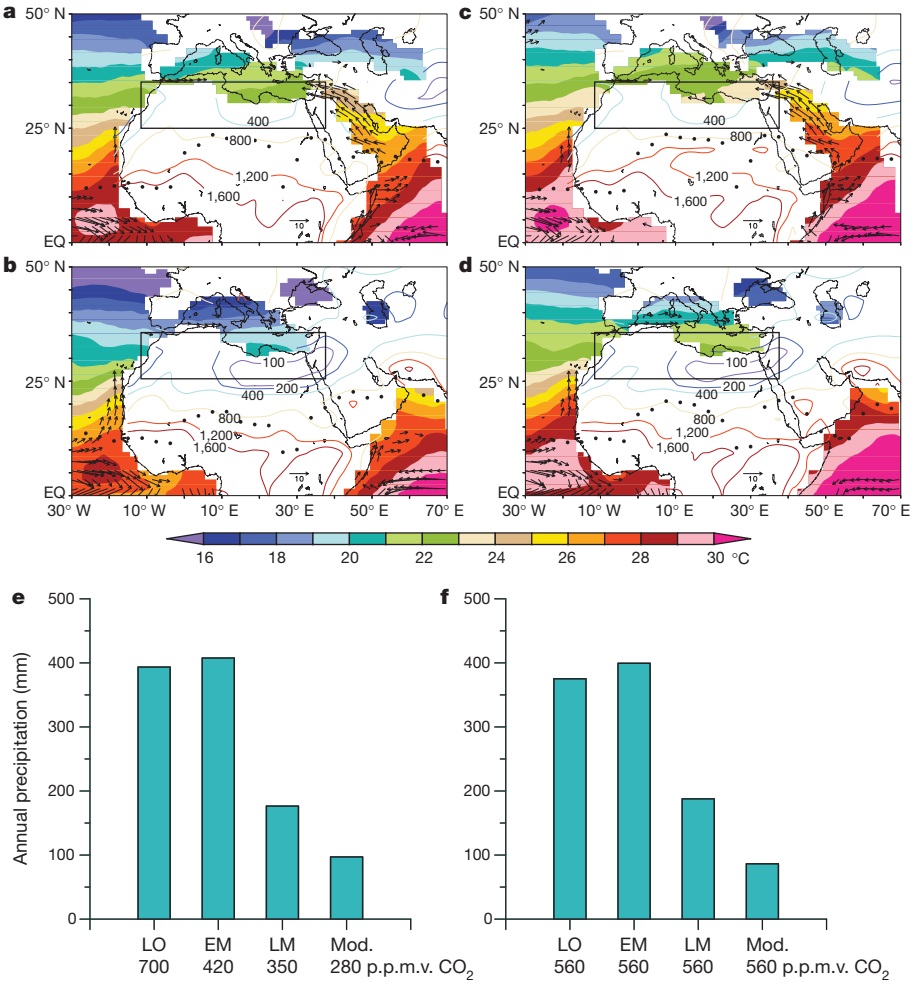


Figure 2 | North African precipitation from coupled experiments. Annual precipitation (mm, contours), sea surface temperature (°C, shading) and currents (cm s^{-1}, arrows) simulated in the experiments with varying CO_2 (**a**, **b**) and at fixed CO_2 (**c**, **d**). **a**, Early Miocene with CO_2 at 420 p.p.m. by volume (p.p.m.v.); **b**, Late Miocene with 350 p.p.m.v. CO_2; **c**, Early Miocene with 560 p.p.m.v. CO_2; **d**, Late Miocene with 560 p.p.m.v. CO_2. The black dots show the positions of climatological mean ITCZ calculated according to ref. 6. EQ, Equator. **e**, **f**, Annual precipitation (mm) averaged over north Africa (between 25° N and 35° N, 12° W and 38° E; black boxes in **a**–**d**) in the Late Oligocene (LO), Early Miocene (EM), Late Miocene (LM) and modern experiments (Mod.) with the indicated atmospheric CO_2 levels.

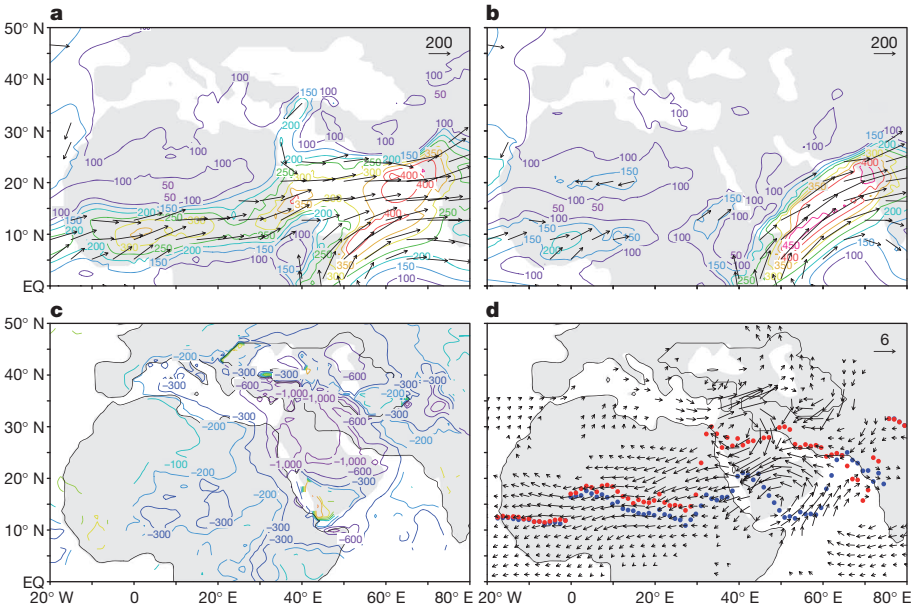


Figure 3 | Changing moisture transport and precipitation due to Tethys shrinkage. **a**, **b**, Column-integrated moisture transport with a large Tethys, representing the Early Miocene (**a**), and modern land–sea distribution, representing the Late Miocene (**b**). The contour lines show magnitude ($>100\,kg\,m^{-1}\,s^{-1}$ displayed) and the arrows show the direction ($>150\,kg\,m^{-1}\,s^{-1}$ displayed). **c**, **d**, Changes in annual precipitation (mm, contours) (**c**) and summer 850-hPa winds ($m\,s^{-1}$, arrows) (**d**) due to Tethys shrinkage. Only changes that are significant at the 95% confidence level (two-tailed unequal *t*-test) are shown. The red (a large Tethys) and blue (modern land–sea distribution) dots show the positions of climatological mean ITCZ[6].

the most notable event affecting land–sea distribution was the shrinkage of the Tethys Sea[25–27], a body of water comprising the proto-Mediterranean, West Tethys and Paratethys seas that is the origin of the modern Mediterranean, Black and Caspian seas (Fig. 1). In the Late Oligocene and Early Miocene, the Tethys was large, with the West Tethys seaway connecting the proto-Mediterranean Sea to the Indian Ocean (Fig. 1a, b)[25–27]. The Afro-Arabian and Eurasian plates collided during the Middle Miocene, closing the West Tethys seaway[25]. Finally, during the Tortonian stage (~7–11 Myr ago) of the Late Miocene, the West Tethys was replaced by the Arabian peninsula, and the Paratethys became much smaller (Fig. 1c)[27].

The shrinkage of the Tethys drastically weakens the ASM and causes a reorganization of atmospheric moisture transport leading to a significant reduction in north African precipitation, as shown in the high-resolution CAM4 simulations (Fig. 3). When the Tethys is large, large heat fluxes from the ocean surface input energy to the atmospheric column above, which exports the energy via a thermally direct circulation. There is a continuous transport path bringing moisture from the tropical Atlantic across all of north Africa in the ASM (Fig. 3a). Replacing the Tethys with land leads to a substantial decrease in surface heat fluxes as well as in the net energy in the atmospheric column (Extended Data Fig. 7). This energy deficit causes anomalous low-level atmospheric divergence[18] that weakens the ASM and is accompanied by a southward retreat of the climatological mean Intertropical Convergence Zone in eastern north Africa in both the atmosphere-only experiment (Fig. 3d) and the coupled experiment (Fig. 2). As a result, less moisture is transported from the tropical Atlantic and precipitation is reduced in north Africa.

In addition to causing a change in mean climate, the Tethys shrinkage also increases north African climate sensitivity to orbital forcing (Fig. 4 and Extended Data Figs 8 and 9). As described above, when Northern Hemisphere summer insolation is stronger (such as at 6 kyr ago), the amplified land–sea contrast sharpens the pressure gradient between the north African continent and the tropical Atlantic and strengthens the ASM. However, a given insolation change causes a weaker atmospheric response (near-surface winds and precipitation shown in Fig. 4) with a large Tethys than with a modern land–sea distribution. These experiments thus predict a change in how the precession cycle drives climatic and hydrological variations depending on the land–sea distribution of the region.

Of the various stages in the shrinking of the Tethys Sea, the replacement of the West Tethys by the Arabian peninsula during the Tortonian stage (~7–11 Myr ago) of the Late Miocene seems most important for the onset of Saharan aridity (see Methods and Extended Data Fig. 4). Before the Tortonian, north Africa had a semiarid climate (Fig. 1a, b) and was not very sensitive to orbital changes (Fig. 4a), implying that the region was generally covered with vegetation and lakes, much as it was

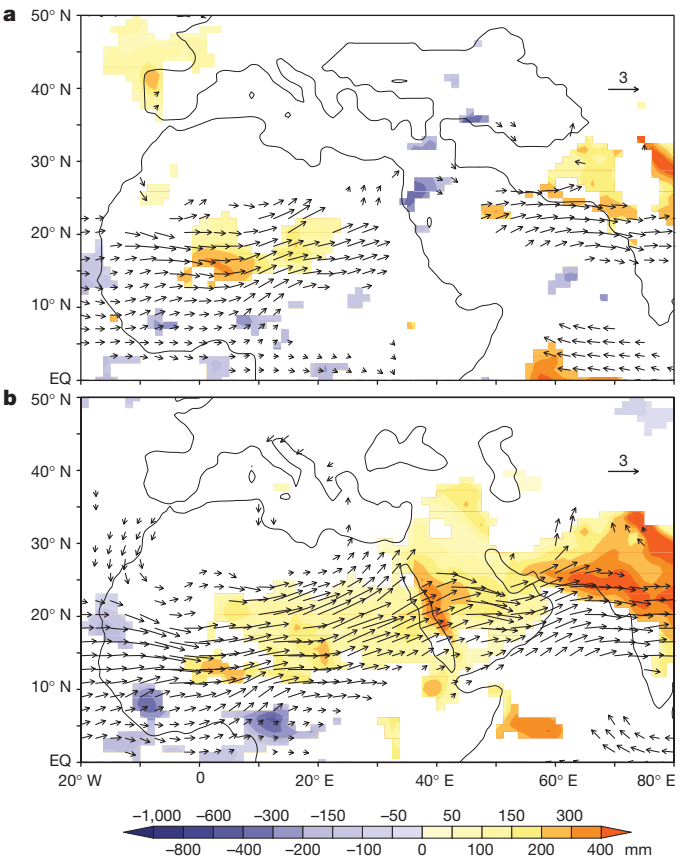


Figure 4 | Climate response to increased summer insolation in the Northern Hemisphere before and after Tethys shrinkage. Response in annual precipitation (mm, shading) and 850-hPa summer winds ($m\,s^{-1}$, arrows) to a change in orbital parameters from today to 6 kyr ago (higher minus lower Northern Hemisphere summer insolation) for continental configuration with a large Tethys (**a**) and a modern land–sea distribution (**b**). Only changes that are significant at the 95% confidence level (two-tailed unequal *t*-test) are shown.

during the Quaternary African Humid Periods[7,8]. After the Tortonian, north Africa became more arid (Fig. 1c, d) on average and desert conditions prevailed, but with periodic fluctuations to a green Sahara in response to orbital forcing, as expected from our experiments (Fig. 4b). The arid–humid fluctuations are evident from the Tortonian to the present[8,17], paced by the ~20-kyr precession cycles during the Late Miocene and Pliocene[2], over which glacial–interglacial cycles are superimposed during the Quaternary[2].

Previous studies have linked the Tethys shrinkage with climate changes further east, in particular those associated with an enhancement of the South Asian summer monsoon (SASM)[28]. Geological evidence shows increased upwelling in the Indian Ocean at ~7–8 Myr ago[29], an indicator for wind changes due to a stronger SASM. Additionally, modelling experiments show that central Eurasian summer temperatures increase in response to the Tethys shrinkage, which would also enhance the monsoon circulation[28]. Our study further supports this idea. When the West Tethys Sea is replaced by the Arabian peninsula, the resulting changes in the atmospheric energy balance produce a near-surface cyclonic anomaly that strengthens the western branch of the SASM (the Somali jet; Fig. 3).

Here we have identified the pivotal role of the shrinkage of the Tethys Sea during the Tortonian stage for north African aridification. Modelling experiments show that the altered land–sea distribution affects the region's monsoon circulations and enhances sensitivity to orbital forcing, permitting the appearance of oscillating arid (desert) and humid (green Sahara) periods paced by precession. These results indicate that the Tortonian stage was critical in the shift from a permanently vegetated Sahara to a desert exhibiting arid–humid oscillations at orbital timescales. This climate shift could be responsible for the pronounced shifts in flora and fauna that occurred during the Late Miocene, for example the replacement of C_3-photosynthesizing plants by C_4 plants[4,12], the demise of Tethyan laurel forests[13] and the split of the endemic African mammalian order Macroscelidea into two species[14], all of which have been ascribed to the widespread aridification of north Africa and Asia. Moreover, the marked change to modern-like climate conditions in the region is likely to have affected the emergence of early hominins in Africa, notably the Sahara[15]. The heightened sensitivity to orbital forcing after the Tortonian has intriguing links to mounting evidence that changes in precipitation due to precession were fundamental to the evolution and dispersal of hominins in north Africa[16].

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 28 March; accepted 21 July 2014.

1. Kroepelin, S. Revisiting the age of the Sahara desert. *Science* **312,** 1138–1139 (2006).
2. deMenocal, P. Plio-Pleistocene African climate. *Science* **270,** 53–59 (1995).
3. Ruddiman, W. F. *et al.* Late Miocene to Pleistocene evolution of climate in Africa and the low-latitude Atlantic: overview of Leg 108 results. *Proc. ODP Sci. Results* **108,** 463–484 (1989).
4. Feakins, S. J. *et al.* Northeast African vegetation change over 12 m.y. *Geology* **41,** 295–298 (2013).
5. Kutzbach, J. E. & Liu, Z. Response of the African monsoon to orbital forcing and ocean feedbacks in the middle Holocene. *Science* **278,** 440–443 (1997).
6. Braconnot, P. *et al.* Results of PMIP2 coupled simulations of the Mid-Holocene and Last Glacial Maximum, part 2. Feedbacks with emphasis on the location of the ITCZ and mid- and high latitudes heat budget. *Clim. Past* **3,** 279–296 10.5194/cp-3–279–2007 (2007).
7. Tierney, J. E. & deMenocal, P. B. Abrupt shifts in Horn of Africa hydroclimate since the Last Glacial Maximum. *Science* **342,** 843–846 (2013).
8. Larrasoana, J. C., Roberts, A. P. & Rohling, E. J. Dynamics of green Sahara periods and their role in hominin evolution. *PLoS ONE* **8,** e76514 (2013).
9. Schuster, M. *et al.* The age of the Sahara desert. *Science* **311,** 821 (2006).
10. Zhang, Z.-S. *et al.* Pre-industrial and mid-Pliocene simulations with NorESM-L. *Geosci. Model Dev.* **5,** 523–533 (2012).
11. Neale, R. B. *et al. Description of the NCAR Community Atmosphere Model (CAM 4.0)* (National Center for Atmospheric Research, 2010).
12. Edwards, E. J. *et al.* The origins of C_4 grasslands: integrating evolutionary and ecosystem science. *Science* **328,** 587–591 (2010).
13. Rodríguez-Sánchez, F. & Arroyo, J. In *Climate Change, Ecology and Systematics* (eds Hodkinson, T. R. *et al.*) 280–303 (Cambridge Univ. Press, 2011).
14. Douady, C. J., Catzeflis, F., Raman, J., Springer, M. S. & Stanhope, M. J. The Sahara as a vicariant agent, and the role of Miocene climatic events, in the diversification of the mammalian order Macroscelidea (elephant shrews). *Proc. Natl Acad. Sci. USA* **100,** 8325–8330 (2003).
15. Brunet, M. *et al.* A new hominid from the Upper Miocene of Chad, Central Africa. *Nature* **418,** 145–151 (2002).
16. Shultz, S. & Maslin, M. A. Early human speciation, brain expansion and dispersal influenced by African climate pulses. *PLoS ONE* **8,** e76750 (2013).
17. Schuster, M. *et al.* Chad Basin: paleoenvironments of the Sahara since the Late Miocene. *C. R. Geosci.* **341,** 603–611 (2009).
18. Chou, C. & Neelin, J. D. Mechanisms limiting the northward extent of the northern summer convection zones. *J. Clim.* **16,** 406–425 (2003).
19. Maslin, M. A. & Christensen, B. Tectonics, orbital forcing, global climate change, and human evolution in Africa. *J. Hum. Evol.* **53,** 443–464 (2007).
20. Yin, A. & Harrison, T. M. Geologic evolution of the Himalayan–Tibetan orogeny. *Annu. Rev. Earth Planet. Sci.* **28,** 211–280 (2000).
21. Tapponnier, P. *et al.* Oblique stepwise rise and growth of the Tibet Plateau. *Science* **294,** 1671–1677 (2001).
22. Sepulchre, P. *et al.* Tectonic uplift and eastern African aridification. *Science* **313,** 1419–1423 (2006).
23. Wu, G. *et al.* Thermal controls on the Asian summer monsoon. *Sci. Rep.* **2,** 404 (2012).
24. Boos, W. R. & Kuang, Z. Dominant control of the South Asian monsoon by orographic insulation versus plateau heating. *Nature* **463,** 218–222 (2010).
25. Rögl, F. Mediterranean and Paratethys. Facts and hypotheses of an Oligocene to Miocene paleogeography (short overview). *Geol. Carpathica* **50,** 339–349 (1999).
26. Scotese, C. R. *Digital Paleogeographic Map Archive on CD-ROM* (PALEOMAP Project, 2001).
27. Barrier, E. & Vrielynck, B. *Palaeotectonic Maps of the Middle East* (Commission for the Geological Map of the World, 2008).
28. Ramstein, G., Fluteau, F., Besse, J. & Joussaume, S. Effect of orogeny, plate motion and land–sea distribution on Eurasian climate change over the past 30 million years. *Nature* **386,** 788–795 (1997).
29. Gupta, A. K., Singh, R. K., Joseph, S. & Thomas, E. Indian Ocean high-productivity event (10–8 Ma): linked to global cooling or to the initiation of the Indian monsoons? *Geology* **32,** 753–756 (2004).

Acknowledgements We thank N. Caud for her help in preparing the paper, and F. Guy for discussions on hominin emergence and evolution. This study was jointly supported by the Strategic and Special Frontier Project of Science and Technology of the Chinese Academy of Sciences (grant no. XDA05080803); the National 973 Program of China (grant no. 2010CB950102); the Earth System Modelling (ESM) project financed by Statoil, Norway; the Dynamics of Past Warm Climates (DYNAWARM) project financed by the Centre for Climate Dynamics (SKD) at the Bjerknes Centre; and the Aurora mobility program France–Norway financed by the Research Council of Norway.

Author Contributions Z.Z. designed and performed the simulations, and wrote the draft of the paper. G.R., M.S. and C.C. linked the Saharan aridification to the emergence and evolution of early hominins. C.L. and Q.Y. checked the analyses in atmospheric dynamics. All authors contributed to discussion of the results and writing of the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Z.Z (zhongshi.zhang@uni.no).

LETTER

doi:10.1038/nature13728

Spreading continents kick-started plate tectonics

Patrice F. Rey[1], Nicolas Coltice[2,3] & Nicolas Flament[1]

Stresses acting on cold, thick and negatively buoyant oceanic lithosphere are thought to be crucial to the initiation of subduction and the operation of plate tectonics[1,2], which characterizes the present-day geodynamics of the Earth. Because the Earth's interior was hotter in the Archaean eon, the oceanic crust may have been thicker, thereby making the oceanic lithosphere more buoyant than at present[3], and whether subduction and plate tectonics occurred during this time is ambiguous, both in the geological record and in geodynamic models[4]. Here we show that because the oceanic crust was thick and buoyant[5], early continents may have produced intra-lithospheric gravitational stresses large enough to drive their gravitational spreading, to initiate subduction at their margins and to trigger episodes of subduction. Our model predicts the co-occurrence of deep to progressively shallower mafic volcanics and arc magmatism within continents in a self-consistent geodynamic framework, explaining the enigmatic multimodal volcanism and tectonic record of Archaean cratons[6]. Moreover, our model predicts a petrological stratification and tectonic structure of the sub-continental lithospheric mantle, two predictions that are consistent with xenolith[5] and seismic studies, respectively, and consistent with the existence of a mid-lithospheric seismic discontinuity[7]. The slow gravitational collapse of early continents could have kick-started transient episodes of plate tectonics until, as the Earth's interior cooled and oceanic lithosphere became heavier, plate tectonics became self-sustaining.

Present-day plate tectonics is primarily driven by the negative buoyancy of cold subducting plates. Petrological and geochemical proxies of subduction preserved in early continents point to subduction-like processes already operating before 3 billion years (Gyr) ago[8,9] and perhaps as early as 4.1 Gyr ago[10]. However, they are not unequivocal, and geodynamic modelling suggests that the thicker basaltic crust produced by partial melting of a hotter Archaean or Hadean mantle would have had increased lithospheric buoyancy and inhibited subduction[3,4]. Mantle convection under a stagnant lid with extensive volcanism could therefore have preceded the onset of subduction[11]. In this scenario, it is classically assumed that the transition from stagnant-lid regime to mobile-lid regime and the onset of plate tectonics require that convective stresses overcame the strength of the stagnant lid[12] at some stage in the Archaean.

On the modern Earth, gravitational stresses due to continental buoyancy can contribute to the initiation of subduction[2,13]. The role of continental gravitational stresses as a driver of Archaean lithospheric deformation has been emphasized[14,15]; however, their potential to initiate subduction has been overlooked. Studies of xenoliths from Archaean cratons show that the early continental crust was underlain by a thick (~200 km) lithospheric mantle, moderately to strongly depleted and therefore buoyant[5]. A common model for the formation of early continental lithosphere invokes partial melting in mantle plumes, leading to magnesium-rich mantle residues (for example, refractory harzburgites and dunites) under thick basaltic plateaux[5,16,17]. Partial melting of these thick basaltic crusts, at depths >40 km, further differentiates the crust into tonalite–trondjemite–granodiorite (TTG) and restitic material[16,18].

First-order calculations show that the horizontal gravitational force acting between a continent 200 km thick and adjacent oceanic lithosphere is of the order of 10^{13} N m^{-1} (see Extended Data Fig. 1), comparable to that of present-day tectonic forces driving orogenesis[1]. To explore the tectonic impact of a thick and buoyant continent surrounded by a stagnant lithospheric lid, we produced a series of two-dimensional thermo-mechanical numerical models of the top 700 km of the Earth, using temperature-dependent densities and visco-plastic rheologies that depend on temperature, melt fraction and depletion, stress and strain rate (see Methods). The initial temperature field is the horizontally averaged temperature profile of a stagnant-lid convection calculation for a mantle ~200 K hotter than at present (Fig. 1A, a and Extended Data Fig. 2). The absence of lateral temperature gradients ensures that no convective stresses act on the lid, allowing us to isolate the dynamic effects of the continent. A buoyant and stiff continent 225 km thick (strongly depleted mantle root 170 km thick overlain by felsic crust 40 km thick; see Fig. 1B, a) is inserted within the lid, on the left side of the domain to exploit the symmetry of the problem (Fig. 1A, a). A mafic crust 15 km thick covers the whole system (Fig. 1A, a), consistent with the common occurrence of thick greenstone covers on continents, as well as thick basaltic crust on the oceanic lid[3].

Our numerical solutions show that the presence of a buoyant continent imparts a horizontal force large enough to induce a long period (~50–150 Myr) of slow collapse of the whole continental lithosphere (Fig. 1 and Extended Data Fig. 3), in agreement with the dynamics of spreading for gravity currents[19]. Hence, a continent of larger volume leads to larger gravitational power and faster collapse. Because of lateral spreading of the continent, the adjacent lithospheric lid is slowly pushed under its margin (Fig. 1A, b and Extended Data Fig. 3A, a). For gravitational stress lower than the yield stress of the oceanic lid, thickening of the margin of the lid is slow, and viscous drips (that is, Rayleigh–Taylor instabilities) detach from its base (Extended Data Fig. 3A, a and b). These instabilities, typical of stagnant-lid convection[20], mitigate the thermal thickening of the lid.

When gravitational stresses overcome the yield stress of the lithospheric lid, subduction is initiated (Fig. 1A, b and c). Depending on the half-width of the continent and its density contrast with the adjacent oceanic lid (that is, its gravitational power) three situations can arise: first, subduction initiates and stalls (Extended Data Fig. 3b); second, the slab detaches and the lid stabilizes (Fig. 1A, d and e and Extended Data Fig. 3c); or third, recurrent detachment of the slab continues until recycling of the oceanic lid is completed, followed by stabilization (Extended Data Fig. 3d). When the slab reaches a depth of ~200 km, slab pull contributes to drive subduction and rapid rollback of the subducting lid, which in turn promotes lithospheric boudinage and continental rifting (Fig. 1A, c and Extended Data Fig. 3C, b–d). Through spreading and thinning of the continent, its base rises from 225 km to ~75 km deep on average, and shallower between lithospheric boudins (Figs 1A, d and 2b and Extended Data Fig. 3C, c and d and D, b–f). This triggers an episode of deep (~150 km) to shallow (<100 km) decompression melting and progressive depletion of the ambient fertile mantle (Figs 2 and 3b). Harzburgites of the continental mantle are too refractory to melt on decompression. Polybaric melting of fertile mantle produces a basalt cover 6 km thick and a mantle residue ~75 km thick with an average depletion of 7.5% (Figs 2 and 3c). The bulk of depletion occurs during a volcanic episode lasting up to ~13 Myr, although partial melting persists for up to 45 Myr (Fig. 2b).

[1]Earthbyte Research Group, School of Geosciences, The University of Sydney, Sydney NSW 2006, Australia. [2]Laboratoire de Géologie de Lyon, UMR 5276 CNRS, Université Lyon 1, Ecole Normale Supérieure de Lyon, 69622 Villeurbanne Cedex, France. [3]Institut Universitaire de France, 75005 Paris, France.

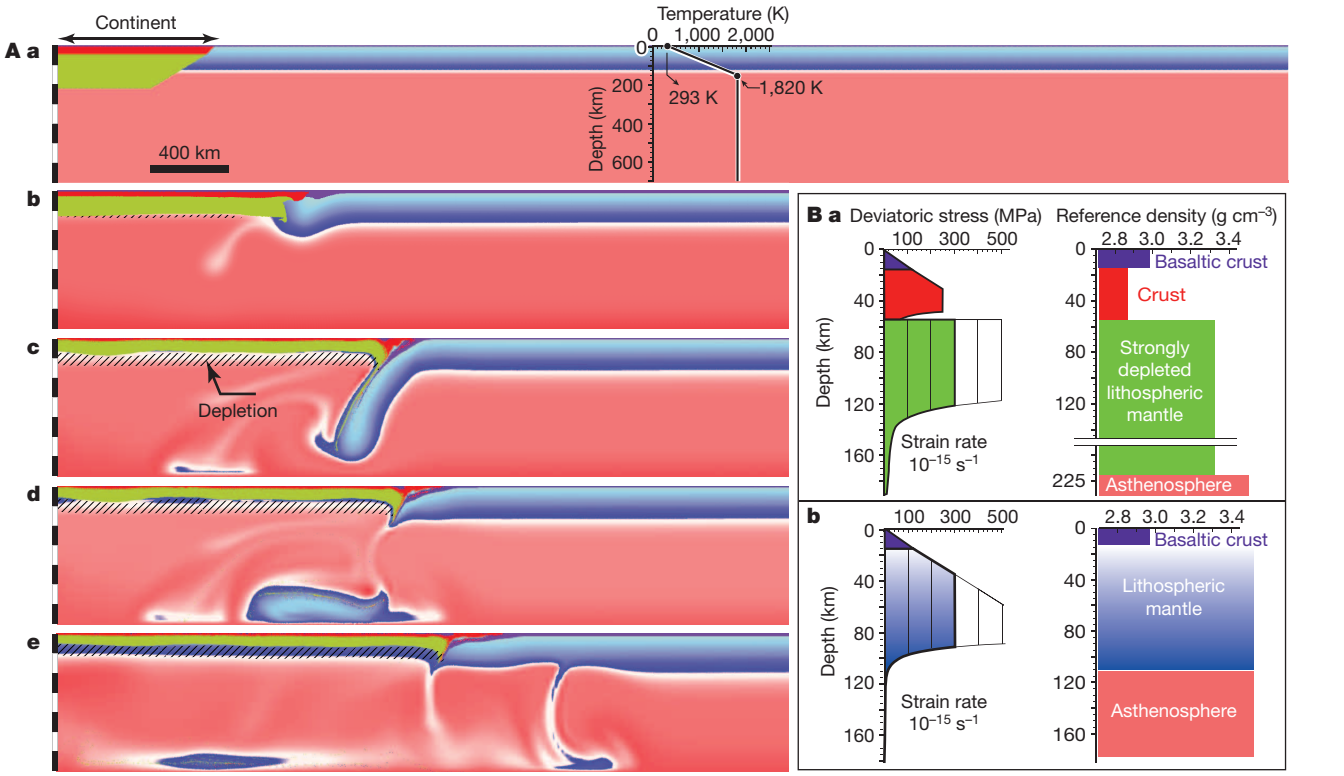


Figure 1 | Numerical solution of an example of continent collapse leading to subduction. **A**, **a**, Modelling setup (0 Myr). **b**–**e**, Computed snapshots for a box 700 km deep and 6,300 km long including a continent 225 km thick with a half-width of 800 km. **b**, 46.7 Myr; **c**, 55.3 Myr; **d**, 57.2 Myr; **e**, 123.8 Myr. All mantle rocks have a limiting yield stress of 300 MPa. Mantle cooler than 1,620 K is in blue (darker blue is hotter); mantle hotter than 1,620 K is in pink (darker pink is hotter). Regions of depletion due to partial melting of ambient fertile mantle are hatched. **B**, Compositional structure, reference densities and reference rheological profile for the continent (**a**) and for the adjacent lithospheric lid (**b**). This numerical solution documents the long phase of slow continental spreading leading to the initiation of a slab (**A**, **b** and **c**). Once the slab has reached a depth of ~200 km, slab pull contributes to drive subduction, rollback and continental boudinage (**A**, **c**) (in some experiments boudinage leads to rifting) and slab detachment (**A**, **d**). In this experiment the detachment of the slab is followed by a long period of thermal relaxation and stabilization during which the thickness of the continent increases through cooling and incorporation of the moderately depleted mantle (**A**, **e**).

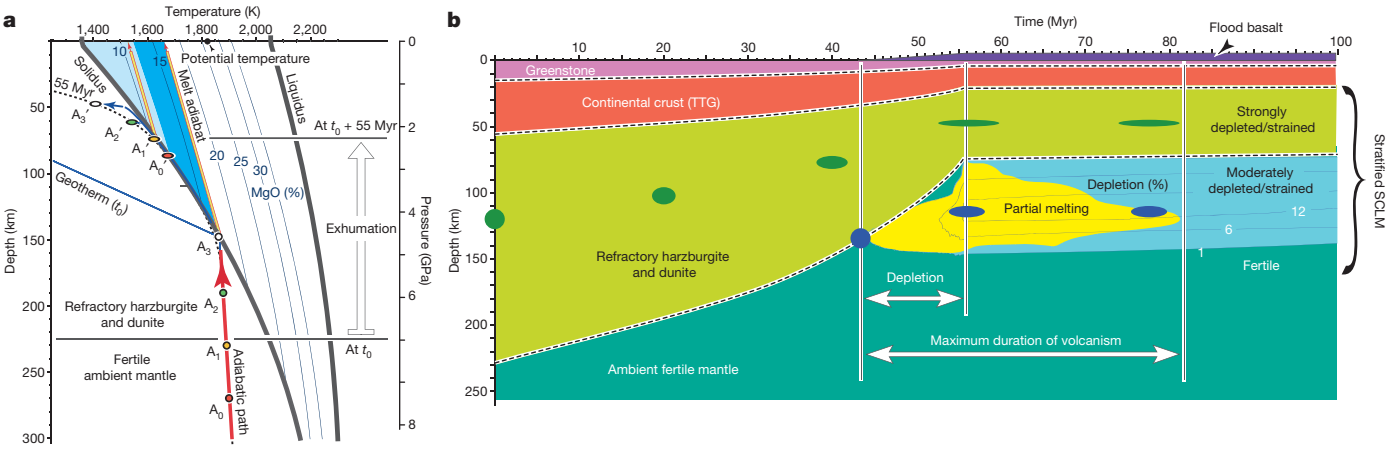


Figure 2 | Development of layering of the continental lithosphere through thinning and progressive accretion of moderately depleted mantle. **a**, Points located at depths A_0 to A_3 before spreading are exhumed during spreading to locations A_0' to A_3', following the blue pressure–temperature–time path. The geotherm intersects the solidus, and the temperature in the partially molten column remains close to the solidus because latent heat is continuously extracted with the melt once melt fraction has reached 1%. Melt is extracted from various depths following the melt adiabat (yellow paths). The region between the solidus and liquidus of the hydrous fertile mantle is mapped for MgO content (see Methods). The deeper part of the column produces komatiitic basalts (dark blue shading), whereas partial melting at pressures <3 GPa produces tholeiitic basalts (pale blue shading). **b**, Temporal evolution of the laterally averaged depletion (blue), partial melting (yellow) and density interfaces (thick dashed lines). As spreading and thinning proceed, pure shear fabrics (shown as finite strain ellipses) develop in the refractory mantle and in the moderately depleted mantle, which records a shorter strain history. The base of the partially molten column remains close to ~150 km, whereas its top progressively rises from ~150 km at the beginning of partial melting (at ~44 Myr) to an average of ~75 km below the surface (at ~55 Myr). From 55 Myr, spreading slows down and progressive cooling reduces the amount of melt, until partial melting stops at ~82 Myr. This results in the progressive chemical and structural stratification of the lithospheric mantle.

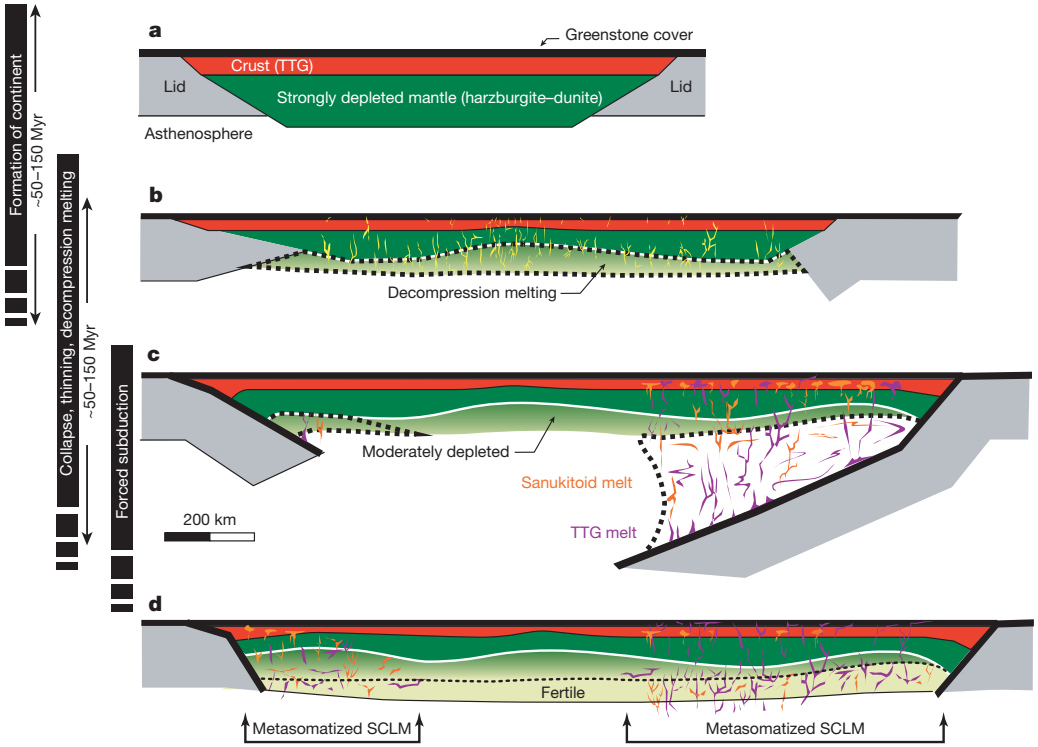


Figure 3 | Proposed model for the co-evolution of cratonic crust and sub-continental lithospheric mantle. Integration of the results of our numerical experiments with petrological data supports a model linking the formation of continents to the initiation of subduction at their margins, through the process of continent collapse. This model predicts the layering of the sub-continental lithospheric mantle (SCLM), polybaric and multimodal volcanism recorded in greenstone cover, and the metasomatism of the SCLM. **a**, Partial melting in the deep mantle leads to the formation of an oceanic plateau that differentiates into a continent. The residue of mantle melting forms the strongly depleted harzburgite root of the continent, whereas the deeper part of the basaltic crust differentiates by partial melting into TTG. **b**, As the continent grows in thickness and length, excess gravitational potential energy drives its collapse and the shortening of the adjacent oceanic lid (in grey). During collapse and thinning of the continent, decompression melting of the fertile ambient mantle and extraction of deep (komatiitic basalt) to shallow (tholeiite) melts (yellow) contribute to the growth of the greenstone cover, and to the formation of a moderately depleted mantle layer. **c**, As a result of the horizontal push of the collapsing continent, the thickened margin of the oceanic lid subducts underneath the continental margin. Partial melting of the thick, eclogitized oceanic crust produces TTG melts (purple), which metasomatise both the mantle wedge and the lithospheric mantle. Melting of the hydrated and metasomatized mantle wedge produces calc-alkaline to sanukitoid melts (orange). **d**, After detachment of the slab, and once the gravitational power of the continent is too small to deform its surrounding, the continent thickens through thermal relaxation and cooling, first incorporating the layer of moderately depleted mantle and then a layer of unmelted fertile mantle.

In the last ~25 Myr of melting, the sub-continental mantle cools until melting stops (Figs 1A, e and 2). Decompression melting allows the spreading continent to maintain a minimum chemical thickness of at least 140–150 km. After the melting phase, conductive cooling results in the thermal thickening and strengthening of a chemically stratified cratonic lithosphere (Figs 1A, e, 2 and 3d). Over the whole process, the buoyancy of the continent decreases, subduction stops and a stagnant lid regime is re-established (Fig. 1A, e).

Trade-offs between yield stress, gravitational stress and continental volume determine the initiation of subduction in our models (Extended Data Fig. 3b, d). For a yield stress of 150–300 MPa, consistent with recent estimates of rheological parameters of the lithospheric mantle[21,22], increasing the width of the continent favours the initiation of subduction (Extended Data Fig. 4). These results suggest an increasing potential for subduction as continental area increased over time.

Not only do our models confirm important results from previous studies, but they also provide innovative explanations for key attributes of Archaean cratons. Both transient subduction and dripping styles are consistent with previous models of Archaean[23] and modern[13] subduction. As observed in previous work, the length of the detached segments increases with the yield stress of the lid[23]. Our models confirm that a combination of the buoyancy and high viscosity of Archaean sub-continental lithospheric mantle, and large plastic strain weakening, prevent the recycling of the sub-continental lithospheric mantle, which explains its longevity[24]. Our models account for the average thickness (~6 km) and duration of volcanism (~10–50 Myr) of greenstone covers, and predict polybaric (5–2 GPa) partial melting, involving deep (garnet-bearing) to shallow sources (Fig. 2). The MgO content of basaltic melts produced at these pressures ranges between 11% and 17% (Fig. 2a), consistent with the komatiitic basalts (12–18% MgO) typical of Archaean greenstones[17,25]. In our model, tholeiitic basalts (6–12% MgO), abundant in Archean greenstones[25], can be produced in regions of continental necking where partial melting can occur at pressures <3 GPa. Figure 2a shows that to account for the formation of komatiites (MgO > 18%) our model would simply require a mantle potential temperature greater than 1,820 K. When subduction starts, arc volcanism is expected at convergent margins while continental extension and rifting still operate (Fig. 1A, c). Our model can therefore explain the metasomatism and production of sanukitoid melts through the migration of younger TTG melts generated by the partial melting of subducting eclogitized basaltic crust (Fig. 3c).

On modern Earth, mantle plumes mostly occur away from subduction zones[26]. Hence, the eruption duration and sequential interlayering of komatiite, tholeiite, calc-alkaline and felsic volcanics, ubiquitous in Archaean greenstone belts[25,27,28], frequently attributed to repeated interaction between mantle plumes and subduction zones over hundreds of millions of years[6], remains enigmatic. However, our model predicts this co-occurrence of deep (up to 150 km) to shallow (<100 km) mafic volcanics (Fig. 2b) and arc magmatism, in a self-consistent geodynamic framework.

Moreover, our model predicts a progressive chemical stratification of the sub-continental lithospheric mantle concomitant with that of the

continental crust and growth of the greenstone cover. This is consistent with the strong geochemical layering of cratonic mantle inferred by geochemical and petrological studies of mantle xenoliths[5]. Pure shear stretching and thinning during the collapse promotes development of a subhorizontal litho-tectonic fabric in the refractory harzburgite and dunite, and to a smaller extent in the accreted moderately depleted mantle (Fig. 2b). The predicted litho-tectonic layering can explain the seismic mid-lithospheric discontinuity at about 100 km depth observed within cratons[29]. This discontinuity[7] could correspond to the sharp transition predicted by our model between the strongly stretched, strongly depleted primary root of the continent and the moderately stretched, moderately depleted-to-fertile mantle accreted through cooling (Fig. 3).

We propose that the collapse of early continents was a key process in Archaean geodynamics, resulting in the concomitant structuration of the mantle root and the crust of cratons. This process would have kick-started transient episodes of plate tectonics, until plate tectonics became self-sustaining through the increasing continental area[30] and the decreasing buoyancy of oceanic plates[3].

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 26 March; accepted 30 July 2014.

1. Bott, M. H. P. Modelling the plate-driving mechanism. *J. Geol. Soc. Lond.* **150,** 941–951 (1993).
2. Faccenna, C., Giardini, D., Davy, P. & Argentieri, A. Initiation of subduction at Atlantic-type margins: insights from laboratory experiments. *J. Geophys. Res.* **104,** 2749–2766 (1999).
3. Sleep, N. H. & Windley, B. F. Archean plate tectonics: constraints and inferences. *J. Geol.* **90,** 363–379 (1982).
4. van Hunen, J. & Moyen, J. F. Archean subduction: fact or fiction? *Annu. Rev. Earth Planet. Sci.* **40,** 195–219 (2012).
5. Griffin, W. L., O'Reilly, S. Y., Afonso, J.-C. & Begg, G. C. The composition and evolution of lithospheric mantle: a re-evaluation and its tectonic implications. *J. Petrol.* **50,** 1185–1204 (2009).
6. Wyman, D. A., Kerrich, R. & Polat, A. Assembly of Archean cratonic mantle lithosphere and crust: plume–arc interaction in the Abitibi–Wawa subduction accretion complex. *Precambr. Res.* **115,** 37–62 (2002).
7. Yuan, H. & Romanowicz, B. Lithospheric layering in the North American craton. *Nature* **466,** 1063–1068 (2010).
8. Martin, H., Smithies, R. H., Rapp, R., Moyen, J. F. & Champion, D. An overview of adakite, tonalite–trondhjemite–granodiorite (TTG), and sanukitoid: relationships and some implications for crustal evolution. *Lithos* **79,** 1–24 (2005).
9. Polat, A., Hofmann, A. W. & Rosing, M. T. Boninite-like volcanic rocks in the 3.7–3.8 Ga Isua greenstone belt, West Greenland: geochemical evidence for intra-oceanic subduction zone processes in the early Earth. *Chem. Geol.* **184,** 231–254 (2002).
10. Harrison, T. M. *et al.* Heterogeneous Hadean hafnium: evidence of continental crust at 4.4 to 4.5 Ga. *Science* **310,** 1947–1950 (2005).
11. van Thienen, P., Vlaar, N. J. & van den Berg, A. J. Assessment of the cooling capacity of plate tectonics and flood volcanism in the evolution of Earth, Mars and Venus. *Phys. Earth Planet. Inter.* **150,** 287–315 (2005).
12. Moresi, L., Dufour, F. & Muhlhaus, H.-B. A Lagrangian integration point finite element method for large deformation modeling of viscoelastic geomaterials. *J. Comput. Phys.* **184,** 476–497 (2003).
13. Nikolaeva, K., Gerya, T. V. & Marques, F. O. Subduction initiation at passive margins: numerical modeling. *J. Geophys. Res.* **115,** B03406 (2010).
14. Bailey, R. Gravity-driven continental overflow and Archaean tectonics. *Nature* **398,** 413–415 (1999).
15. Rey, P. F. & Coltice, N. Neoarchean lithospheric strengthening and the coupling of the Earth's geochemical reservoirs. *Geology* **36,** 635–638 (2008).
16. Bédard, J. H. A catalytic delamination-driven model for coupled genesis of Archaean crust and sub-continental lithospheric mantle. *Geochim. Cosmochim. Acta* **70,** 1188–1214 (2006).
17. Arndt, N. T., Coltice, N., Helmstaedt, H. & Michel, G. Origin of Archean subcontinental lithospheric mantle: some petrological constraints. *Lithos* **109,** 61–71 (2009).
18. Zhang, C. *et al.* Constraints from experimental melting of amphibolite on the depth of formation of garnet-rich restites, and implications for models of Early Archean crustal growth. *Precambr. Res.* **231,** 206–217 (2013).
19. Huppert, H. E. Propagation of two-dimensional and axisymmetric viscous gravity currents over a rigid horizontal surface. *J. Fluid Mech.* **121,** 43–58 (1982).
20. Moresi, L. & Solomatov, V. Mantle convection with a brittle lithosphere: thoughts on the global tectonic styles of the Earth and Venus. *Geophys. J. Int.* **133,** 669–682 (1998).
21. Demouchy, S., Tommasi, A., Ballaran, T. B. & Cordier, P. Low strength of Earth's uppermost mantle inferred from tri-axial deformation experiments on dry olivine crystals. *Phys. Earth Planet. Inter.* **220,** 37–49 (2013).
22. Zhong, S. & Watts, A. B. Lithospheric deformation induced by loading of the Hawaiian Islands and its implications for mantle rheology. *J. Geophys. Res.* **118,** 1–24 (2013).
23. van Hunen, J. & van den Berg, A. P. Plate tectonics on the early Earth: limitations imposed by strength and buoyancy of subducted lithosphere. *Lithos* **103,** 217–235 (2008).
24. Lenardic, A., Moresi, L.-N. & Mühlaus, H. Longevity and stability of cratonic lithosphere: insights from numerical simulations of coupled mantle convection and continental tectonics. *J. Geophys. Res.* **108,** 2303 (2003).
25. Sproule, R. A., Lesher, C. M., Ayer, J., Thurston, P. C. & Herzberg, C. T. Spatial and temporal variations in the geochemistry of komatiitic rocks in the Abitibi greenstone belt. *Precambr. Res.* **115,** 153–186 (2002).
26. Cazenave, A., Souriau, A. & Dominh, K. Global coupling of Earth surface topography with hotspots, geoid and mantle heterogeneities. *Nature* **340,** 54–57 (1989).
27. Ayer, J. *et al.* Evolution of the southern Abitibi greenstone belt based on U–Pb geochronology: autochthonous volcanic construction followed by plutonism, regional deformation and sedimentation. *Precambr. Res.* **115,** 63–95 (2002).
28. Bateman, R., Costa, S., Swe, T. & Lambert, D. Archaean mafic magmatism in the Kalgoorlie area of the Yilgarn Craton, Western Australia: a geochemical and Nd isotopic study of the petrogenetic and tectonic evolution of a greenstone belt. *Precambr. Res.* **108,** 75–112 (2001).
29. Kind, R., Yuan, X. & Kumar, P. Seismic receiver functions and the lithosphere–asthenosphere boundary. *Tectonophysics* **536–537,** 25–43 (2012).
30. Rolf, T. & Tackley, P. J. Focussing of stress by continents in 3D spherical mantle convection with self-consistent plate tectonics. *Geophys. Res. Lett.* **38,** L18301 (2011).

Acknowledgements We thank W. L. Griffin for comments on the manuscript. P.F.R. acknowledges the assistance of resources provided at the NCI National Facility systems at the Australian National University through the National Computational Merit Allocation Scheme supported by the Australian Government. N.C. was supported by the Institut Universitaire de France, and the European Research Council (ERC) within the framework of the SP2-Ideas Program ERC-2013-CoG, under ERC grant agreement no. 617588. N.C. acknowledges discussions with B. Romanowicz and B. Tauzin. N.F. was supported by Statoil ASA.

Author Contributions P.F.R. conceived the study. P.F.R. and N.C. performed numerical experiments. P.F.R., N.C. and N.F. interpreted the results. P.F.R., N.C. and N.F. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.F.R. (patrice.rey@sydney.edu.au).

LETTER

doi:10.1038/nature13673

Ancient human genomes suggest three ancestral populations for present-day Europeans

A list of authors and their affiliations appears at the end of the paper

We sequenced the genomes of a ~7,000-year-old farmer from Germany and eight ~8,000-year-old hunter-gatherers from Luxembourg and Sweden. We analysed these and other ancient genomes[1–4] with 2,345 contemporary humans to show that most present-day Europeans derive from at least three highly differentiated populations: west European hunter-gatherers, who contributed ancestry to all Europeans but not to Near Easterners; ancient north Eurasians related to Upper Palaeolithic Siberians[3], who contributed to both Europeans and Near Easterners; and early European farmers, who were mainly of Near Eastern origin but also harboured west European hunter-gatherer related ancestry. We model these populations' deep relationships and show that early European farmers had ~44% ancestry from a 'basal Eurasian' population that split before the diversification of other non-African lineages.

Near Eastern migrants from Anatolia and the Levant are known to have played a major role in the introduction of agriculture to Europe, as ancient DNA indicates that early European farmers were distinct from European hunter-gatherers[4,5] and close to present-day Near Easterners[4,6]. However, modelling present-day Europeans as a mixture of these two ancestral populations[4] does not account for the fact that Europeans are also admixed with a population related to Native Americans[7,8]. To clarify the prehistory of Europe, we sequenced nine ancient genomes (Fig. 1 and Extended Data Fig. 1): 'Stuttgart' (19-fold coverage), a ~7,000-year-old skeleton found in Germany in the context of artefacts from the first widespread farming culture of central Europe, the Linearbandkeramik; 'Loschbour' (22-fold), an ~8,000-year-old skeleton from the Loschbour rock shelter in Luxembourg, discovered in the context of hunter-gatherer artefacts (Supplementary Information sections 1 and 2); and seven ~8,000-year-old samples (0.01–2.4-fold) from a hunter-gatherer burial in Motala, Sweden (the highest coverage individual was 'Motala12').

Sequence reads from all samples revealed >20% C→T and G→A deamination-derived mismatches at the ends of the molecules that are characteristic of ancient DNA[9,10] (Supplementary Information section 3). We estimate nuclear contamination rates to be 0.3% for Stuttgart and 0.4% for Loschbour (Supplementary Information section 3), and mitochondrial (mtDNA) contamination rates to be 0.3% for Stuttgart, 0.4% for Loschbour, and 0.01–5% for the Motala individuals (Supplementary Information section 3). Stuttgart has mtDNA haplogroup T2, typical of Neolithic Europeans[11], and Loschbour and all Motala individuals have the U5 or U2 haplogroups, typical of hunter-gatherers[5,9] (Supplementary Information section 4). Stuttgart is female, whereas Loschbour and five Motala individuals are male (Supplementary Information section 5) and belong to Y-chromosome haplogroup I, suggesting that this was common in pre-agricultural Europeans (Supplementary Information section 5).

We carried out large-scale sequencing of libraries prepared with uracil DNA glycosylase (UDG), which removes deaminated cytosines, thus reducing errors arising from ancient DNA damage (Supplementary Information section 3). The ancient individuals had indistinguishable levels of Neanderthal ancestry when compared to each other (~2%) and to present-day Eurasians (Supplementary Information section 6). The heterozygosity of Stuttgart (0.00074) is at the high end of present-day Europeans, whereas that of Loschbour (0.00048) is lower than in any present human populations (Supplementary Information section 2); this must reflect a strong bottleneck in Loschbour's ancestors, as the genetic data show that he was not recently inbred (Extended Data Fig. 2). High copy numbers for the salivary amylase gene (*AMY1*) have been associated with a high starch diet[12]; our ancient genomes are consistent with the direction of this observation in that the Stuttgart farmer had the highest number of copies (16), whereas the ancient hunter-gatherers La Braña (from Iberia)[2], Motala12, and Loschbour had lower numbers (5, 6 and 13, respectively) (Supplementary Information section 7). We caution, however, that copy count in Loschbour is at the high end of present-day humans, showing that high copy counts of *AMY1* cannot be accounted for entirely by selection since the switch to agriculture. Both Loschbour and Stuttgart had dark hair (>99% probability); and Loschbour, like La Braña and Motala12, probably had blue or light coloured eyes (>75%) whereas Stuttgart probably had brown eyes (>99% probability) (Supplementary Information section 8). Neither Loschbour nor La Braña carries the skin-lightening allele in *SLC24A5* that is homozygous in Stuttgart and nearly fixed in Europeans today[2], but Motala12 carries at least one copy of the derived allele, showing that this allele was present in Europe before the advent of agriculture.

We compared the ancient genomes to 2,345 present-day humans from 203 populations genotyped at 594,924 autosomal single nucleotide polymorphisms (SNPs) with the Human Origins array[8] (Supplementary Information section 9) (Extended Data Table 1). We used ADMIXTURE[13] to identify 59 'west Eurasian' populations that cluster with Europe and the Near East (Supplementary Information section 9 and Extended Data Fig. 3). Principal component analysis (PCA)[14] (Supplementary Information section 10) (Fig. 2) indicates a discontinuity between the Near East and Europe, with each showing north–south clines bridged only by a few populations of mainly Mediterranean origin. We projected[15] the newly sequenced and previously published[1–4] ancient genomes onto the first two principal components (PCs) (Fig. 2). Upper Palaeolithic hunter-gatherers[3] from Siberia like the MA1 (Mal'ta) individual project at the northern end of the PCA, suggesting an 'ancient north Eurasian' (ANE) meta-population. European hunter-gatherers from Spain[2], Luxembourg, and Sweden[4] fall beyond present-day Europeans in the direction of European differentiation from the Near East, and form a 'west European hunter-gatherer' (WHG) cluster including Loschbour and La Braña[2], and a 'Scandinavian hunter-gatherer' (SHG) cluster including the Motala individuals and ~5,000-year-old hunter-gatherers from the Pitted Ware Culture[4]. An 'early European farmer' (EEF) cluster includes Stuttgart, the ~5,300-year-old Tyrolean Iceman[1] and a ~5,000-year-old Swedish farmer[4].

Patterns observed in PCA may be affected by sample composition (Supplementary Information section 10) and their interpretation in terms of admixture events is not straightforward, so we rely on formal analysis of *f* statistics[8] to document mixture of at least three source populations in the ancestry of present Europeans. We began by computing all possible statistics of the form $f_3(Test; Ref_1, Ref_2)$ (Supplementary Information section 11), which if significantly negative show unambiguously[8] that *Test* is admixed between populations anciently related to Ref_1 and Ref_2 (we choose Ref_1 and Ref_2 from 5 ancient and 192 present populations). The lowest f_3 statistics for Europeans are negative (93% are > 4 standard errors below 0), with most showing strong support for at least one

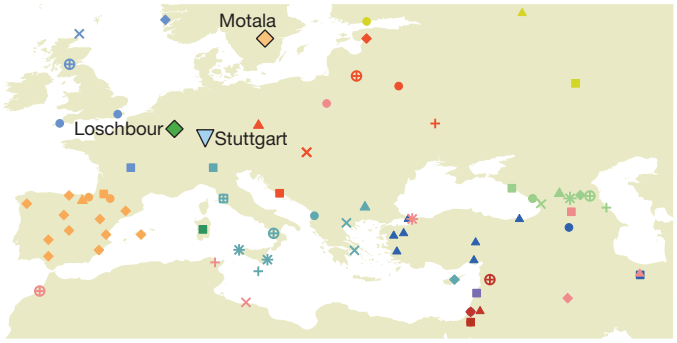


Figure 1 | Map of west Eurasian populations. Geographical locations of analysed samples, with colour coding matching the PCA (Fig. 2). We show all sampling locations for each population, which results in multiple points for some (for example, Spain).

ancient individual being one of the references (Supplementary Information section 11). Europeans almost always have their lowest f_3 with either (EEF, ANE) or (WHG, Near East) (Supplementary Information section 11, Table 1 and Extended Data Table 1), which would not be expected if there were just two ancient sources of ancestry (in which case the best references for all Europeans would be similar). The lowest f_3 statistic for Near Easterners always takes Stuttgart as one of the reference populations, consistent with a Near Eastern origin for Stuttgart's ancestors (Table 1). We also computed the statistic f_4(*Test,* Stuttgart; MA1, Chimp), which measures whether MA1 shares more alleles with a *Test* population or with Stuttgart. This statistic is significantly positive (Extended Data Fig. 4 and Extended Data Table 1) if *Test* is nearly any present-day West Eurasian population, showing that MA1-related ancestry has increased since the time of early farmers like Stuttgart (the same statistic using Native Americans instead of MA1 has the same sign but is smaller in magnitude (Extended Data Fig. 5), indicating that MA1 is a better surrogate than the Native Americans who were first used to document ANE ancestry in Europe[7,8]). The analogous statistic f_4(*Test,* Stuttgart; Loschbour, Chimp) is nearly always positive in Europeans and negative in Near Easterners, indicating that Europeans have more ancestry from populations related to Loschbour than do Near Easterners (Extended Data Fig. 4 and Extended Data Table 1). Extended Data Table 2 documents the robustness of key f_4 statistics by recomputing them using transversion polymorphisms not affected by ancient DNA damage, and also using whole-genome sequencing data not affected by SNP ascertainment bias. Extended Data Fig. 6 shows the geographic gradients in the degree of allele sharing of present-day West Eurasians (as measured by f_4 statistics) with Stuttgart (EEF), Loschbour (WHG) and MA1 (ANE).

To determine the minimum number of source populations needed to explain the data for many European populations taken together, we studied the matrix of all possible statistics of the form $f_4(Test_{base}, Test_i; O_{base}, O_j)$ (Supplementary Information section 12). $Test_{base}$ is a reference European population, $Test_i$ is the set of all other European *Test* populations, O_{base} is a reference outgroup, and O_j is the set of other outgroups (ancient DNA samples, Onge, Karitiana, and Mbuti). The rank of the (i, j) matrix reflects the minimum number of sources that contributed to the *Test* populations[16,17]. For a pool of individuals from 23 *Test* populations representing most present-day European groups, this analysis rejects descent from just two sources ($P < 10^{-12}$ by a Hotelling *t*-test[17]). However, three source populations are consistent with the data after excluding the Spanish who have evidence for African admixture[18–20] ($P = 0.019$, not significant after multiple-hypothesis correction), consistent with the results from ADMIXTURE (Supplementary Information section 9), PCA (Fig. 2 and Supplementary Information section 10) and *f* statistics (Extended Data Table 1, Extended Data Fig. 6, Supplementary Information sections 11 and 12). We caution that the finding of three sources could be consistent with a larger number of mixture events. Moreover, the source

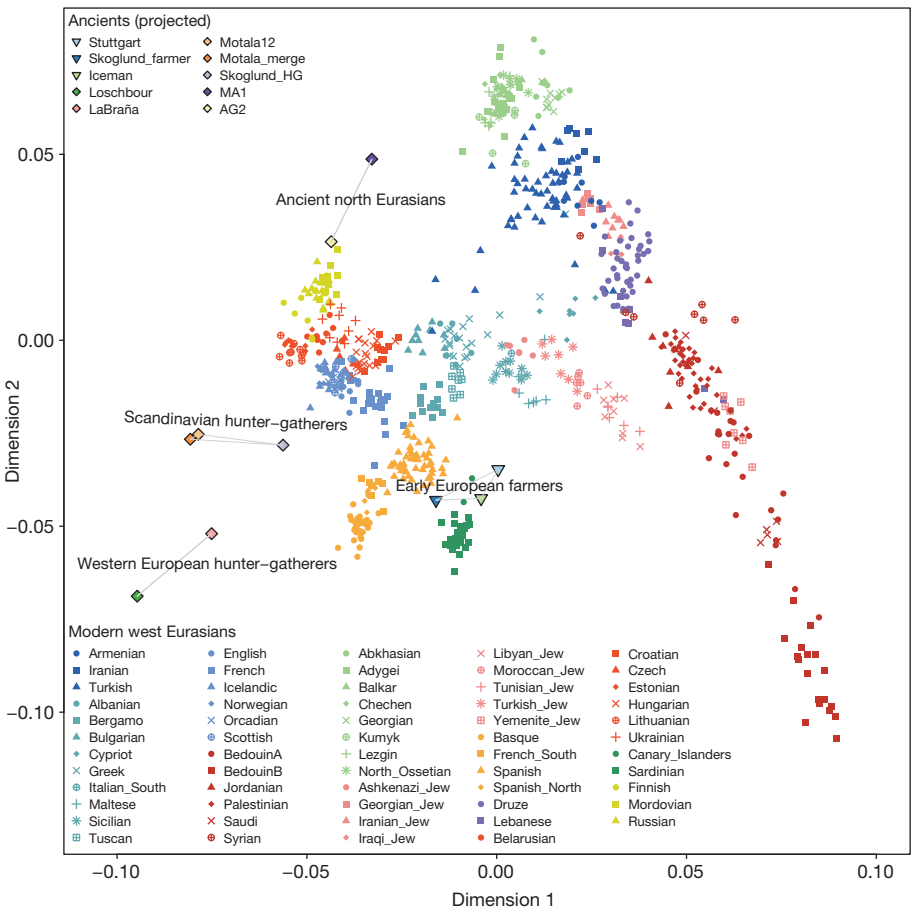


Figure 2 | Principal Component Analysis. PCA on all present-day west Eurasians, with ancient samples projected. European hunter-gatherers fall beyond present-day Europeans in the direction of European differentiation from the Near East. Stuttgart clusters with other Neolithic Europeans and present-day Sardinians. MA1 falls outside the variation of present-day west Eurasians in the direction of southern–northern differentiation along dimension 2.

Table 1 | Lowest f_3 statistics for each west Eurasian population

Ref_1	Ref_2	Target for which these two references give the lowest $f_3(X; Ref_1, Ref_2)$
WHG	EEF	Sardinian***
WHG	Near East	Basque, Belarusian, Czech, English, Estonian, Finnish, French_South, Icelandic, Lithuanian, Mordovian, Norwegian, Orcadian, Scottish, Spanish, Spanish_North, Ukrainian
WHG	Siberian	Russian
EEF	ANE	Abkhasian***, Albanian, Ashkenazi_Jew****, Bergamo, Bulgarian, Chechen****, Croatian, Cypriot****, Druze**, French, Greek, Hungarian, Lezgin, Maltese, Sicilian, Turkish_Jew, Tuscan
EEF	Native American	Adygei, Balkar, Iranian, Kumyk, North_Ossetian, Turkish
EEF	African	BedouinA, BedouinB†, Jordanian, Lebanese, Libyan_Jew, Moroccan_Jew, Palestinian, Saudi****, Syrian, Tunisian_Jew***, Yemenite_Jew***
EEF	South Asian	Armenian, Georgian****, Georgian_Jew*, Iranian_Jew***, Iraqi_Jew***

WHG = Loschbour or LaBraña; EEF = Stuttgart; ANE = MA1; Native American = Piapoco; African = Esan, Gambian, or Kgalagadi; South Asian = GujaratiC or Vishwabrahmin. Statistics are negative with $Z < -4$ unless otherwise noted: †(positive) or *, **, ***, ****, to indicate Z less than 0, −1, −2, and −3, respectively. The complete list of statistics can be found in Extended Data Table 1.

populations may themselves have been mixed. Indeed, the positive f_4 (Stuttgart, *Test*; Loschbour, Chimp) statistics obtained when *Test* is Near Eastern (Extended Data Table 1) imply that the EEF had some WHG-related ancestry, which was greater than 0% and as high as 45% (Supplementary Information section 13).

We used the ADMIXTUREGRAPH software[8,15] to fit a model (a tree structure augmented by admixture events) to the data, exploring models relating the three ancient populations (Stuttgart, Loschbour, and MA1) to two eastern non-Africans (Onge and Karitiana) and sub-Saharan Africans (Mbuti). We found no models that fit the data with 0 or 1 admixture events, but did find a model that fit with 2 admixture events (Supplementary Information section 14). The successful model (Fig. 3) confirms the existence of MA1-related admixture in Native Americans[3], but includes the novel inference that Stuttgart is partially (44 ± 10%) derived from a lineage that split before the separation of eastern non-Africans from the common ancestor of WHG and ANE. The existence of such basal Eurasian admixture into Stuttgart provides a simple explanation for our finding that diverse eastern non-African populations share significantly more alleles with ancient European and Upper Palaeolithic Siberian hunter-gatherers than with Stuttgart (that is, f_4(*Eastern non-African*, Chimp; *Hunter-gatherer*, Stuttgart) is significantly positive), but that hunter-gatherers appear to be equally related to most eastern groups (Supplementary Information section 14). We verified the robustness of the model by reanalysing the data using the unsupervised MixMapper[7] (Supplementary Information section 15) and TreeMix[21] software (Supplementary Information section 16), which both identified the same admixture events. The ANE–WHG split must have occurred > 24,000 years ago (as it must predate the age of MA1 (ref. 3)), and the WHG and Eastern non-African split must have occurred > 40,000 years ago (as it must predate the Tianyuan[22] individual from China which clusters with Asians to the exclusion of Europeans). The basal Eurasian split must be even older, and might be related to early settlement of the Levant[23] or Arabia[24,25] before the diversification of most Eurasians, or more recent gene flow from Africa[26]. However, the basal Eurasian population shares much of the genetic drift common to non-African populations after their separation from Africans, and thus does not appear to represent gene flow between sub-Saharan Africans and the ancestors of non-Africans after the out-of-Africa bottleneck (Supplementary Information section 14).

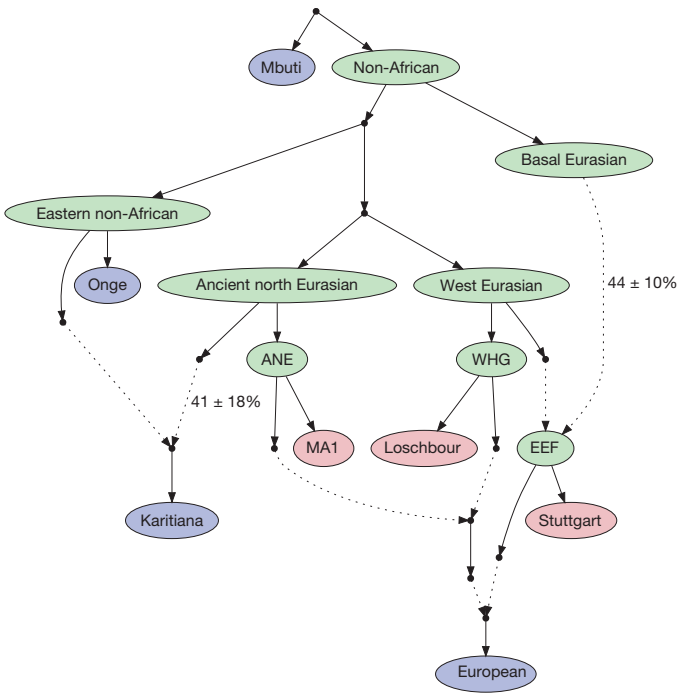


Figure 3 | Modelling the relationship of European to non-European populations. A three-way mixture model that is a fit to the data for many populations. Present-day samples are coloured in blue, ancient in red, and reconstructed ancestral populations in green. Solid lines represent descent without mixture, and dashed lines represent admixture. We print mixture proportions and one standard error for the two mixtures relating the highly divergent ancestral populations. (We do not print the estimate for the 'European' population as it varies depending on the population.)

Fitting present-day Europeans into the model, we find that few populations can be fit as two-way mixtures, but nearly all are compatible with three-way mixtures of ANE–EEF–WHG (Supplementary Information section 14). The mixture proportions from the fitted model (Fig. 4 and Extended Data Table 3) are encouragingly consistent with those obtained from a separate method that relates European populations to diverse outgroups using f_4 statistics, assuming only that MA1 is an unmixed descendent of ANE, Loschbour of WHG, and Stuttgart of EEF (Supplementary Information section 17). We infer that EEF ancestry in Europe today ranges from ~30% in the Baltic region to ~90% in the Mediterranean, consistent with patterns of identity-by-descent (IBD) sharing[27,28] (Supplementary Information section 18) and shared haplotype analysis (chromosome painting)[29] (Supplementary Information section 19) in which Loschbour shares more segments with northern Europeans and Stuttgart with southern Europeans. Southern Europeans inherited their European hunter-gatherer ancestry mostly via EEF ancestors (Extended Data Fig. 6), whereas northern Europeans acquired up to 50% of WHG ancestry above and beyond what they received through their EEF ancestors. Europeans have a larger proportion of WHG than ANE ancestry in general. By contrast, in the Near East there is no detectable WHG ancestry, but up to ~29% ANE in the Caucasus (Supplementary Information section 14). A striking feature of these findings is that ANE ancestry is inferred to be present in nearly all Europeans today (with a maximum of ~20%), but was absent in both farmers and hunter-gatherers from central and western Europe during the Neolithic transition. However, ANE ancestry was not completely absent from the larger European region at that time: we find that it was present in ~8,000-years-old Scandinavian hunter-gatherers, as MA1 shares more alleles with Motala12 (SHG) than with Loschbour, and Motala12 fits as a mixture of 81% WHG and 19% ANE (Supplementary Information section 14).

Two sets of European populations are poor fits for the model. Sicilians, Maltese, and Ashkenazi Jews have EEF estimates of >100%, consistent with their having more Near Eastern ancestry than can be explained via

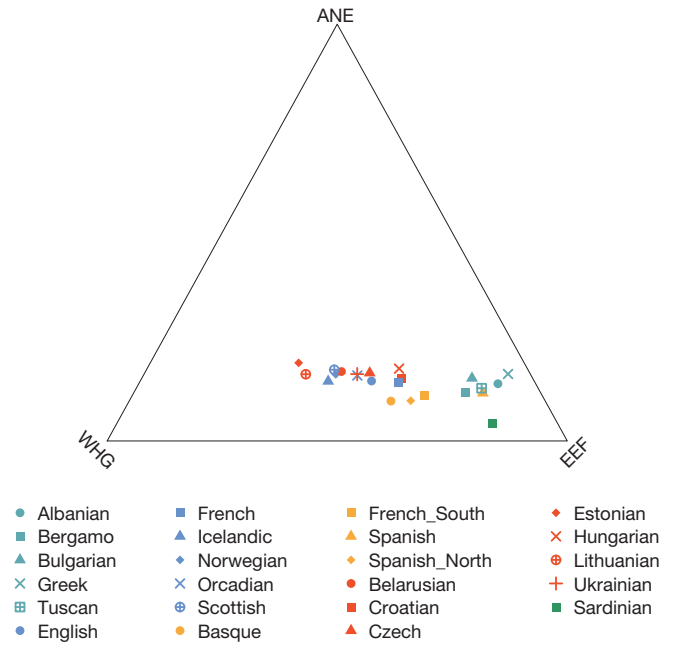


Figure 4 | Estimates of mixture proportions in present-day Europeans. Plot of the proportions of ancestry from each of three inferred ancestral populations (EEF, ANE and WHG).

EEF admixture (Supplementary Information section 17). They also cannot be jointly fit with other Europeans (Supplementary Information section 14), and they fall in the gap between European and Near Easterners in PCA (Fig. 2). Finns, Mordovians and Russians (from the northwest of Russia) also do not fit (Supplementary Information section 14; Extended Data Table 3) due to East Eurasian gene flow into the ancestors of these north-eastern European populations. These populations (and Chuvash and Saami) are more related to east Asians than can be explained by ANE admixture (Extended Data Fig. 7), probably reflecting a separate stream of Siberian gene flow into north-eastern Europe (Supplementary Information section 14).

Several questions will be important to address in future ancient DNA work. One question concerns where and when the Near Eastern farmers mixed with European hunter-gatherers to produce the EEF. A second question concerns how the ancestors of present-day Europeans first acquired their ANE ancestry. Discontinuity in central Europe during the late Neolithic (~4,500 years ago) associated with the appearance of mtDNA types absent in earlier farmers and hunter-gatherers[30] raises the possibility that ANE ancestry may have also appeared at this time. Finally, it will be important to study ancient genome sequences from the Near East to provide insights into the history of the basal Eurasians.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 23 December 2013; accepted 11 July 2014.

1. Keller, A. *et al.* New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nature Commun.* **3,** 698 (2012).
2. Olalde, I. *et al.* Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature* **507,** 225–228 (2014).
3. Raghavan, M. *et al.* Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* **505,** 87–91 (2014).
4. Skoglund, P. *et al.* Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science* **336,** 466–469 (2012).
5. Bramanti, B. *et al.* Genetic discontinuity between local hunter-gatherers and Central Europe's first farmers. *Science* **326,** 137–140 (2009).
6. Haak, W. *et al.* Ancient DNA from European early Neolithic farmers reveals their Near Eastern affinities. *PLoS Biol.* **8,** e1000536 (2010).
7. Lipson, M. *et al.* Efficient moment-based inference of admixture parameters and sources of gene flow. *Mol. Biol. Evol.* **30,** 1788–1802 (2013).
8. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192,** 1065–1093 (2012).
9. Krause, J. *et al.* A complete mtDNA genome of an early modern human from Kostenki, Russia. *Curr. Biol.* **20,** 231–236 (2010).
10. Sawyer, S., Krause, J., Guschanski, K., Savolainen, V. & Pääbo, S. Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS ONE* **7,** e34131 (2012).
11. Haak, W. *et al.* Ancient DNA from the first European farmers in 7500-year-old Neolithic sites. *Science* **310,** 1016–1018 (2005).
12. Perry, G. H. *et al.* Diet and the evolution of human amylase gene copy number variation. *Nature Genet.* **39,** 1256–1260 (2007).
13. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19,** 1655–1664 (2009).
14. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2,** e190 (2006).
15. Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian population history. *Nature* **461,** 489–494 (2009).
16. Moorjani, P. *et al.* Genetic evidence for recent population mixture in India. *Am. J. Hum. Genet.* **93,** 422–438 (2013).
17. Reich, D. *et al.* Reconstructing Native American population history. *Nature* **488,** 370–374 (2012).
18. Botigué, L. R. *et al.* Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. *Proc. Natl Acad. Sci. USA* **110,** 11791–11796 (2013).
19. Cerezo, M. *et al.* Reconstructing ancient mitochondrial DNA links between Africa and Europe. *Genome Res.* **22,** 821–826 (2012).
20. Moorjani, P. *et al.* The history of African gene flow into southern Europeans, Levantines, and Jews. *PLoS Genet.* **7,** e1001373 (2011).
21. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8,** e1002967 (2012).
22. Fu, Q. *et al.* DNA analysis of an early modern human from Tianyuan Cave, China. *Proc. Natl Acad. Sci. USA* **110,** 2223–2227 (2013).
23. Bar-Yosef, O. *The Chronology of the Middle Paleolithic of the Levant* 39–56 (Plenum Press, 1998).
24. Armitage, S. J. *et al.* The southern route "out of Africa": evidence for an early expansion of modern humans into Arabia. *Science* **331,** 453–456 (2011).
25. Rose, J. I. *et al.* The Nubian Complex of Dhofar, Oman: an African middle stone age industry in Southern Arabia. *PLoS ONE* **6,** e28239 (2011).
26. Brace, C. L. *et al.* The questionable contribution of the Neolithic and the Bronze Age to European craniofacial form. *Proc. Natl Acad. Sci. USA* **103,** 242–247 (2006).
27. Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194,** 459–471 (2013).
28. Ralph, P. & Coop, G. The geography of recent genetic ancestry across Europe. *PLoS Biol.* **11,** e1001555 (2013).
29. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8,** e1002453 (2012).
30. Brandt, G. *et al.* Ancient DNA reveals key stages in the formation of central European mitochondrial genetic diversity. *Science* **342,** 257–261 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the 1,615 volunteers from 147 diverse populations who donated DNA samples and whose genetic data are newly reported in this study. We are grateful to C. Beall, N. Bradman, A. Gebremedhin, D. Labuda, M. Nelis and A. Di Rienzo for sharing DNA samples; to D. Weigel, C. Lanz, V. Schünemann, P. Bauer and O. Riess for support and access to DNA sequencing facilities; to P. Johnson for advice on contamination estimation; to G. Hellenthal for help with the ChromoPainter software; and to P. Skoglund for sharing graphics software. We thank K. Nordtvedt for alerting us to newly discovered Y-chromosome SNPs. We downloaded the POPRES data from dbGaP at (http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000145.v4.p2) through dbGaP accession number phs000145.v1.p2. We thank all the volunteers who donated DNA. We thank the staff of the Unità Operativa Complessa di Medicina Trasfusionale, Azienda Ospedaliera Umberto I, Siracusa, Italy for assistance in sample collection; and The National Laboratory for the Genetics of Israeli Populations for facilitating access to DNA. We thank colleagues at the Applied Genomics at the Children's Hospital of Philadelphia, especially H. Hakonarson, C. Kim, K. Thomas, and C. Hou, for genotyping samples on the Human Origins array. J.Kr., A.M. and C.P. are grateful for support from DFG grant number KR 4015/1-1, the Carl-Zeiss Foundation and the Baden Württemberg Foundation. S.P., G.R., Q.F., C.F., K.P., S.C. and J.Ke. acknowledge support from the Presidential Innovation Fund of the Max Planck Society. G.R. was supported by an NSERC fellowship. J.G.S. acknowledges use of the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by NSF grant number OCI-1053575. E.B. and O.B. were supported by RFBR grants 13-06-00670, 13-04-01711, 13-04-90420 and by the Molecular and Cell Biology Program of the Presidium, Russian Academy of Sciences. B.M. was supported by grants OTKA 73430 and 103983. A.Saj. was supported by a Finnish Professorpool (Paulo Foundation) Grant. The Lithuanian sampling was supported by the LITGEN project (VP1-3.1-ŠMM-07-K-01-013), funded by the European Social Fund under the Global Grant Measure. A.S. was supported by Spanish grants SAF2011-26983 and EM 2012/045. O.U. was supported by Ukrainian SFFS grant F53.4/071. S.A.T. was supported by NIH Pioneer Award 8DP1ES022577-04 and NSF HOMINID award BCS-0827436. K.T. was supported by an Indian CSIR Network Project (GENESIS: BSC0121). L.S. was supported by an Indian CSIR Bhatnagar Fellowship. R.V., M.M., J.P. and E.M. were supported by the European Union Regional Development Fund through the Centre of Excellence in Genomics to the Estonian Biocentre and University of Tartu and by an Estonian Basic Research grant SF0270177As08. M.M. was additionally supported by Estonian Science Foundation grant number 8973. J.G.S. and M.S. were supported by NIH grant GM40282. P.H.S. and E.E.E. were supported by NIH grants HG004120 and

HG002385. D.R. and N.P. were supported by NSF HOMINID award BCS-1032255 and NIH grant GM100233. D.R. and E.E.E. are Howard Hughes Medical Institute investigators. This project has been funded in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government. This Research was supported in part by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

Author Contributions B.B., E.E.E., J.Bu., M.S., S.P., J.Ke., D.R. and J.Kr. supervised the study. I.L., N.P., A.M., G.R., S.M., K.K., P.H.S., J.G.S., S.C., M.L., Q.F., H.L., C.dF., K.P., W.H., M.Met., M.Mey. and D.R. analysed genetic data. F.H., E.F., D.D., M.F., J.-M.G., J.W., A.C. and J.Kr. obtained human remains. A.M., C.E., R.Bo., K.I.B., S.S., C.P., N.R. and J.Kr. processed ancient DNA. I.L., N.P., S.N., N.R., G.A., H.A.B., G.Ba., E.B., O.B., R.Ba., G.Be., H.B.-A., J.Be., F.Be., C.M.B., F.Br., G.B.J.B., F.C., M.C., D.E.C.C., D.Cor., L.D., G.vD., S.D., J.-M.D., S.A.F., I.G.R., M.G., M.H., B.M.H., T.H., U.H., A.R.J., S.K.-Y., R.Kh., E.K., R.Ki., T.K., W.K., V.K., A.K., L.L., S.L., T.L., R.W.M., B.M., E.M., J.Mol., J.Mou., K.N., D.N., T.N., L.O., J.P., F.P., O. P., V.R., F.R., I.R., R.R., H.S., A.Saj., A.Sal., E.B.S., A.Tar., D.T., S.T., I.U., O.U., R.Va., M.Vi., M.Vo., C.A.W., L.Y., P.Z., T.Z., C.C., M.G.T., A.R.-L., S.A.T., L.S., K.T., R.Vi., D.Com., R.S., M.Met., S.P. and D.R. assembled the genotyping dataset. I.L., N.P., D.R. and J.Kr. wrote the manuscript with help from all co-authors.

Author Information The aligned sequences are available through the European Nucleotide Archive under accession number PRJEB6272. The fully public version of the Human Origins dataset can be found at (http://genetics.med.harvard.edu/reichlab/Reich_Lab/Datasets.html). The full version of the dataset (including additional samples) is available to researchers who send a signed letter to D.R. indicating that they will abide by specified usage conditions (Supplementary Information section 9). Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.R. (reich@genetics.med.harvard.edu) or J.Kr. (johannes.krause@uni-tuebingen.de).

Iosif Lazaridis[1,2], Nick Patterson[2], Alissa Mittnik[3], Gabriel Renaud[4], Swapan Mallick[1,2], Karola Kirsanow[5], Peter H. Sudmant[6], Joshua G. Schraiber[6,7], Sergi Castellano[4], Mark Lipson[8], Bonnie Berger[2,8], Christos Economou[9], Ruth Bollongino[5], Qiaomei Fu[1,4,10], Kirsten I. Bos[3], Susanne Nordenfelt[1,2], Heng Li[1,2], Cesare de Filippo[4], Kay Prüfer[4], Susanna Sawyer[4], Cosimo Posth[3], Wolfgang Haak[11], Fredrik Hallgren[12], Elin Fornander[12], Nadin Rohland[1,2], Dominique Delsate[13,14], Michael Francken[15], Jean-Michel Guinet[13], Joachim Wahl[16], George Ayodo[17], Hamza A. Babiker[18,19], Graciela Bailliet[20], Elena Balanovska[21], Oleg Balanovsky[21,22], Ramiro Barrantes[23], Gabriel Bedoya[24], Haim Ben-Ami[25], Judit Bene[26], Fouad Berrada[27], Claudio M. Bravi[20], Francesca Brisighelli[28], George B. J. Busby[29,30], Francesco Cali[31], Mikhail Churnosov[32], David E. C. Cole[33], Daniel Corach[34], Larissa Damba[35], George van Driem[36], Stanislav Dryomov[37], Jean-Michel Dugoujon[38], Sardana A. Fedorova[39], Irene Gallego Romero[40], Marina Gubina[35], Michael Hammer[41], Brenna M. Henn[42], Tor Hervig[43], Ugur Hodoglugil[44], Aashish R. Jha[40], Sena Karachanak-Yankova[45], Rita Khusainova[46,47], Elza Khusnutdinova[46,47], Rick Kittles[48], Toomas Kivisild[49], William Klitz[7], Vaidutis Kučinskas[50], Alena Kushniarevich[51], Leila Laredj[52], Sergey Litvinov[46,47,51], Theologos Loukidis[53]†, Robert W. Mahley[54], Béla Melegh[26], Ene Metspalu[55], Julio Molina[56], Joanna Mountain[57], Klemetti Näkkäläjärvi[58], Desislava Nesheva[45], Thomas Nyambo[59], Ludmila Osipova[35], Jüri Parik[55], Fedor Platonov[60], Olga Posukh[35], Valentino Romano[61], Francisco Rothhammer[62,63,64], Igor Rudan[65], Ruslan Ruizbakiev[66]‡, Hovhannes Sahakyan[51,67], Antti Sajantila[68,69], Antonio Salas[70], Elena B. Starikovskaya[37], Ayele Tarekegn[71], Draga Toncheva[45], Shahlo Turdikulova[72], Ingrida Uktveryte[50], Olga Utevska[73], René Vasquez[74,75], Mercedes Villena[74,75], Mikhail Voevoda[35,76,77], Cheryl A. Winkler[78], Levon Yepiskoposyan[67], Pierre Zalloua[79,80], Tatijana Zemunik[81], Alan Cooper[11], Cristian Capelli[29], Mark G. Thomas[53], Andres Ruiz-Linares[53], Sarah A. Tishkoff[82], Lalji Singh[83]†, Kumarasamy Thangaraj[83], Richard Villems[51,55,84], David Comas[85], Rem Sukernik[37], Mait Metspalu[51], Matthias Meyer[4], Evan E. Eichler[6,86], Joachim Burger[5], Montgomery Slatkin[7], Svante Pääbo[4], Janet Kelso[4], David Reich[1,2,87] & Johannes Krause[3,88,89]

[1]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [2]Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. [3]Institute for Archaeological Sciences, University of Tübingen, Tübingen 72074, Germany. [4]Max Planck Institute for Evolutionary Anthropology, Leipzig 04103, Germany. [5]Institute of Anthropology, Johannes Gutenberg University Mainz, Mainz D-55128, Germany. [6]Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA. [7]Department of Integrative Biology, University of California, Berkeley, California 94720-3140, USA. [8]Department of Mathematics and Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [9]Archaeological Research Laboratory, Stockholm University, 114 18, Sweden. [10]Key Laboratory of Vertebrate Evolution and Human Origins of Chinese Academy of Sciences, IVPP, CAS, Beijing 100049, China. [11]Australian Centre for Ancient DNA and Environment Institute, School of Earth and Environmental Sciences, University of Adelaide, Adelaide, South Australia 5005, Australia. [12]The Cultural Heritage Foundation, Västerås 722 12, Sweden. [13]National Museum of Natural History, L-2160, Luxembourg. [14]National Center of Archaeological Research, National Museum of History and Art, L-2345, Luxembourg. [15]Department of Paleoanthropology, Senckenberg Center for Human Evolution and Paleoenvironment, University of Tübingen, Tübingen D-72070, Germany. [16]State Office for Cultural Heritage Management Baden-Württemberg, Osteology, Konstanz D-78467, Germany. [17]Center for Global Health and Child Development, Kisumu 40100, Kenya. [18]Institutes of Evolution, Immunology and Infection Research, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT, UK. [19]Biochemistry Department, Faculty of Medicine, Sultan Qaboos University, Alkhod, Muscat 123, Oman. [20]Laboratorio de Genética Molecular Poblacional, Instituto Multidisciplinario de Biología Celular (IMBICE), CCT-CONICET & CICPBA, La Plata, B1906APO, Argentina. [21]Research Centre for Medical Genetics, Moscow 115478, Russia. [22]Vavilov Institute for General Genetics, Moscow 119991, Russia. [23]Escuela de Biología, Universidad de Costa Rica, San José 2060, Costa Rica. [24]Institute of Biology, Research group GENMOL, Universidad de Antioquia, Medellín, Colombia. [25]Rambam Health Care Campus, Haifa 31096, Israel. [26]Department of Medical Genetics and Szentagothai Research Center, University of Pécs, Pécs H-7624, Hungary. [27]Al Akhawayn University in Ifrane (AUI), School of Science and Engineering, Ifrane 53000, Morocco. [28]Forensic Genetics Laboratory, Institute of Legal Medicine, Università Cattolica del Sacro Cuore, Rome 00168, Italy. [29]Department of Zoology, University of Oxford, Oxford OX1 3PS, UK. [30]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK. [31]Laboratorio di Genetica Molecolare, IRCCS Associazione Oasi Maria SS, Troina 94018, Italy. [32]Belgorod State University, Belgorod 308015, Russia. [33]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario M5G 1L5, Canada. [34]Servicio de Huellas Digitales Genéticas, School of Pharmacy and Biochemistry, Universidad de Buenos Aires, 1113 CABA, Argentina. [35]Institute of Cytology and Genetics, Siberian Branch of Russian Academy of Sciences, Novosibirsk 630090, Russia. [36]Institute of Linguistics, University of Bern, Bern CH-3012, Switzerland. [37]Laboratory of Human Molecular Genetics, Institute of Molecular and Cellular Biology, Russian Academy of Science, Siberian Branch, Novosibirsk 630090, Russia. [38]Anthropologie Moléculaire et Imagerie de Synthèse, CNRS UMR 5288, Université Paul Sabatier Toulouse III, Toulouse 31000, France. [39]North-Eastern Federal University and Yakut Research Center of Complex Medical Problems, Yakutsk 677013, Russia. [40]Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA. [41]ARL Division of Biotechnology, University of Arizona, Tucson, Arizona 85721, USA. [42]Department of Ecology and Evolution, Stony Brook University, Stony Brook, New York 11794, USA. [43]Department of Clinical Science, University of Bergen, Bergen 5021, Norway. [44]NextBio, Illumina, Santa Clara, California 95050, USA. [45]Department of Medical Genetics, National Human Genome Center, Medical University Sofia, Sofia 1431, Bulgaria. [46]Institute of Biochemistry and Genetics, Ufa Research Centre, Russian Academy of Sciences, Ufa 450054, Russia. [47]Department of Genetics and Fundamental Medicine, Bashkir State University, Ufa 450074, Russia. [48]College of Medicine, University of Arizona, Tucson, Arizona 85724, USA. [49]Division of Biological Anthropology, University of Cambridge, Cambridge CB2 1QH, UK. [50]Department of Human and Medical Genetics, Vilnius University, Vilnius LT-08661, Lithuania. [51]Estonian Biocentre, Evolutionary Biology group, Tartu, 51010, Estonia. [52]Translational Medicine and Neurogenetics, Institut de Génétique et de Biologie Moléculaire et Cellulaire, Illkirch 67404, France. [53]Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, UK. [54]Gladstone Institutes, San Francisco, California 94158, USA. [55]Department of Evolutionary Biology, University of Tartu, Tartu 51010, Estonia. [56]Centro de Investigaciones Biomédicas de Guatemala, Ciudad de Guatemala, Guatemala. [57]Research Department, 23andMe, Mountain View, California 94043, USA. [58]Cultural Anthropology Program, University of Oulu, Oulu 90014, Finland. [59]Department of Biochemistry, Muhimbili University of Health and Allied Sciences, Dar es Salaam 65001, Tanzania. [60]Research Institute of Health, North-Eastern Federal University, Yakutsk 677000, Russia. [61]Dipartimento di Fisica e Chimica, Università di Palermo, Palermo 90128, Italy. [62]Instituto de Alta Investigación, Universidad de Tarapacá, Arica 1000000, Chile. [63]Programa de Genética Humana ICBM Facultad de Medicina Universidad de Chile, Santiago 8320000, Chile. [64]Centro de Investigaciones del Hombre en el Desierto, Arica 1000000, Chile. [65]Centre for Population Health Sciences, The University of Edinburgh Medical School, Edinburgh EH8 9AG, UK. [66]Institute of Immunology, Academy of Science, Tashkent 70000, Uzbekistan. [67]Laboratory of Ethnogenomics, Institute of Molecular Biology, National Academy of Sciences of Armenia, Yerevan 0014, Armenia. [68]Department of Forensic Medicine, Hjelt Institute, University of Helsinki, Helsinki 00014, Finland. [69]Institute of Applied Genetics, Department of Molecular and Medical Genetics, University of North Texas Health Science Center, Fort Worth, Texas 76107, USA. [70]Unidade de Xenética, Departamento de Anatomía Patolóxica e Ciencias Forenses, and Instituto de Ciencias Forenses, Grupo de Medicina Xenómica (GMX), Facultade de Medicina, Universidade de Santiago de Compostela, Galcia 15872, Spain. [71]Research Fellow, Henry Stewart Group, Russell House, London WC1A 2HN, UK. [72]Institute of Bioorganic Chemistry Academy of Sciences Republic of Uzbekistan, Tashkent 100125, Uzbekistan. [73]Department of Genetics and Cytology, V. N. Karazin Kharkiv National University, Kharkiv 61077, Ukraine. [74]Instituto Boliviano de Biología de la Altura, Universidad Mayor de San Andrés, 591 2 La Paz, Bolivia. [75]Universidad Autonoma Tomás Frías, Potosí, Bolivia. [76]Institute of Internal Medicine, Siberian Branch of Russian Academy of Medical Sciences, Novosibirsk 630089, Russia. [77]Novosibirsk State University, Novosibirsk 630090, Russia. [78]Basic Research Laboratory, NCI, NIH, Frederick National Laboratory, Leidos Biomedical, Frederick, Maryland 21702, USA. [79]Lebanese American University, School of Medicine, Beirut 13-5053, Lebanon. [80]Harvard School of Public Health, Boston, Massachusetts 02115, USA. [81]Department of Medical Biology, University of Split, School of Medicine, Split 21000, Croatia. [82]Department of Biology and Genetics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [83]CSIR-Centre for Cellular and Molecular Biology, Hyderabad 500 007, India. [84]Estonian Academy of Sciences, Tallinn 10130, Estonia. [85]Institut de Biologia Evolutiva (CSIC-UPF), Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona 08003, Spain. [86]Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA. [87]Howard Hughes Medical Institute, Harvard Medical School, Boston, Massachusetts 02115, USA. [88]Senckenberg Centre for Human Evolution and Palaeoenvironment, University of Tübingen, 72070 Tübingen, Germany. [89]Max Planck Institut für Geschichte und Naturwissenschaften, Jena 07745, Germany. †Present addresses: Amgen, 33 Kazantzaki Str, Ilioupolis 16342, Athens, Greece (T.L.); Banaras Hindu University, Varanasi 221 005, India (L.S.). ‡Deceased.

LETTER

doi:10.1038/nature13727

Lethal aggression in *Pan* is better explained by adaptive strategies than human impacts

Michael L. Wilson[1,2], Christophe Boesch[3], Barbara Fruth[4,5], Takeshi Furuichi[6], Ian C. Gilby[7,8], Chie Hashimoto[6], Catherine L. Hobaiter[9], Gottfried Hohmann[3], Noriko Itoh[10], Kathelijne Koops[11], Julia N. Lloyd[12], Tetsuro Matsuzawa[6,13], John C. Mitani[14], Deus C. Mjungu[15], David Morgan[16], Martin N. Muller[17], Roger Mundry[18], Michio Nakamura[10], Jill Pruetz[19], Anne E. Pusey[7], Julia Riedel[3], Crickette Sanz[20], Anne M. Schel[21], Nicole Simmons[12], Michel Waller[22], David P. Watts[23], Frances White[22], Roman M. Wittig[3], Klaus Zuberbühler[9,24] & Richard W. Wrangham[25]

Observations of chimpanzees (*Pan troglodytes*) and bonobos (*Pan paniscus*) provide valuable comparative data for understanding the significance of conspecific killing. Two kinds of hypothesis have been proposed. Lethal violence is sometimes concluded to be the result of adaptive strategies, such that killers ultimately gain fitness benefits by increasing their access to resources such as food or mates[1–5]. Alternatively, it could be a non-adaptive result of human impacts, such as habitat change or food provisioning[6–9]. To discriminate between these hypotheses we compiled information from 18 chimpanzee communities and 4 bonobo communities studied over five decades. Our data include 152 killings ($n = 58$ observed, 41 inferred, and 53 suspected killings) by chimpanzees in 15 communities and one suspected killing by bonobos. We found that males were the most frequent attackers (92% of participants) and victims (73%); most killings (66%) involved intercommunity attacks; and attackers greatly outnumbered their victims (median 8:1 ratio). Variation in killing rates was unrelated to measures of human impacts. Our results are compatible with previously proposed adaptive explanations for killing by chimpanzees, whereas the human impact hypothesis is not supported.

Substantial variation exists in rates of killing across chimpanzee study sites[2–5,10–12]. The human impact and adaptive strategies hypotheses both seek to explain this variation, but have contrasting predictions, which we test here (Tables 1 and 2). The human impact hypothesis states that killing is an incidental outcome of aggression, exacerbated by human activities such as deforestation, introducing diseases, hunting or providing food. Accordingly, lethal aggression should be high where human disturbance is high[8].

In contrast, the adaptive strategies hypothesis views killing as an evolved tactic by which killers tend to increase their fitness through increased access to territory, food, mates or other benefits[1–5,10–17]. Kin selection[18] and evolutionary game theory[19] yield a set of specific predictions for how benefits and costs should vary with the context, age, sex, and genetic relatedness of the attackers and targets. Lethal aggression occurs within a diverse set of circumstances, but is expected to be most commonly committed by males; directed towards males; directed towards non-kin, particularly members of other groups; and committed when overwhelming numerical superiority reduces the costs of killing. Previous studies have developed and tested these specific hypotheses[2,5,11–17]; the present study represents the first effort to test multiple hypotheses simultaneously with a comprehensive data set. We assembled data from communities of eastern ($n = 12$) and western ($n = 6$) chimpanzees[20] studied over 426 years (median = 21 years; range: 4–53) and from 4 bonobo communities studied for 92 years (median = 21; range: 9–39; Extended Data Fig. 1). We rated each case of killing as observed, inferred, or suspected (see Methods; Extended Data Tables 1–4). To be conservative, we limited our analyses to those rated 'observed' and 'inferred' unless otherwise noted. We examined contrasting predictions relating to overall patterns of killings (Table 1) and variation among communities (Table 2).

Bonobos are consistently found to be less violent than chimpanzees[2,21], and lower rates of killing are reported for western than eastern chimpanzees[2,11]. The human impact hypothesis could in theory ascribe these variations to different levels of disturbance. In contrast, in behavioural ecology, distinct populations are expected to respond to prevailing ecological circumstances through biological evolution and/or phenotypic

Table 1 | Predicted patterns of lethal aggression

Variable	Human impact hypothesis	Adaptive strategies hypothesis
Chimpanzees kill more than bonobos	None	+
Rate of killing over time	+	None
Sex bias: attackers	None	Mainly males
Sex bias: victims	None	Mainly males
Age of victims	None	Mainly young infants (most vulnerable and/or reduce time to mother's next estrus)
Genetic relatedness of attackers and victims	None	Mainly non-relatives (for example, members of other communities)
Numerical asymmetries	None	Victims greatly outnumbered

[1]Department of Anthropology, University of Minnesota, 395 Humphrey Center, 301 19th Avenue South, Minneapolis, Minnesota 55455, USA. [2]Department of Ecology, Evolution and Behavior, University of Minnesota, 1987 Upper Buford Circle, St Paul, Minnesota 55108, USA. [3]Department of Primatology, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany. [4]Division of Neurobiology, Ludwig-Maximilians Universitaet Muenchen, Germany. [5]Centre for Research and Conservation, Royal Zoological Society of Antwerp, Belgium. [6]Primate Research Institute, Kyoto University, 41-2 Kanrin, Inuyama, Aichi 484-8506, Japan. [7]Department of Evolutionary Anthropology, Duke University, 104 Biological Sciences Building, Box 90383, Durham, North Carolina 27708-0680, USA. [8]School of Human Evolution and Social Change, Arizona State University, PO Box 872402, Tempe, Arizona 85287-2402, USA. [9]School of Psychology and Neuroscience, University of St Andrews, Westburn Lane, St Andrews, Fife KY16 9JP, UK. [10]Wildlife Research Center, Kyoto University, 2-24 Tanaka-Sekiden-Cho, Sakyo, Kyoto, Japan. [11]Division of Biological Anthropology, Department of Archaeology & Anthropology, University of Cambridge, Henry Wellcome Building, Fitzwilliam Street, Cambridge CB2 1QH, UK. [12]Zoology Department, Makerere University, P.O. Box 7062, Kampala, Uganda. [13]Japan Monkey Center, 26 Kanrin, Inuyama, Aichi 484-0081, Japan. [14]Department of Anthropology, University of Michigan, 101 West Hall, 1085 S. University Avenue, Ann Arbor, Michigan 48109, USA. [15]Gombe Stream Research Centre, The Jane Goodall Institute – Tanzania, P.O. Box 1182, Kigoma, Tanzania. [16]The Lester E. Fisher Center for the Study and Conservation of Apes, Lincoln Park Zoo, Chicago, Illinois 60614, USA. [17]Department of Anthropology, MSC01-1040, Anthropology 1, University of New Mexico, Albuquerque, New Mexico 87131, USA. [18]Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany. [19]Department of Anthropology, Iowa State University, 324 Curtiss, Ames, Iowa 50011, USA. [20]Department of Anthropology, Washington University in St Louis, Campus Mailbox 1114, One Brookings Drive, St Louis, Missouri 63130, USA. [21]University of York, Department of Psychology, Heslington, York, YO10 5DD, UK. [22]Department of Anthropology, University of Oregon, Eugene, Oregon 97403, USA. [23]Department of Anthropology, Yale University, 10 Sachem Street, New Haven, Connecticut 06511, USA. [24]Université de Neuchâtel, Institut de Biologie, Rue Emile-Argand 11, 2000 Neuchâtel, Switzerland. [25]Department of Human Evolutionary Biology, Harvard University, 11 Divinity Avenue Cambridge, Massachusetts 02138, USA.

Table 2 | Predicted correlates of number of killings per study community

Variable	Human impact hypothesis	Adaptive strategies hypothesis
Provisioning (provisoned)	+	None
Size of protected area, km^2 (area)	−	None
Disturbance rating (disturbance)	+	None
Eastern vs. western chimpanzees (clade)	None	+
Mean number of adult males (males)	None	+
Mean population density (density)	None	+

flexibility. For bonobos and western chimpanzees, ecological factors apparently allow relatively high gregariousness, which reduces the risk of experiencing a lethal attack[2,11]. Our data set covers all major studies of both species of *Pan*, which include sites with and without a history of provisioning, and with high and low levels of human disturbance, a rating estimated independently by each site's director(s) (Methods; Extended Data Figs 1a and 2a).

We documented killings by chimpanzees in 15 of 18 communities (58 observed, 41 inferred, and 53 suspected cases; Extended Data Tables 1–4) (Fig. 1). For bonobos, we documented only a single (suspected) case, which occurred at Lomako, a never-provisioned site with a low disturbance rating. No killings were recorded at other bonobo sites, including one with a history of provisioning and a high disturbance rating (Wamba). Controlling for years of observation, chimpanzees had a higher rate of killing than bonobos; this difference was statistically significant for eastern but not western chimpanzees (Poisson regression: $n = 22$ communities; estimated coefficients ± s.e. for chimpanzees compared to bonobos: $\beta_0 = -4.5 \pm 1.0$; $\beta_{east} = 3.4 \pm 1.0$, $z = 3.3$, $P = 0.0008$; $\beta_{west} = 0.65 \pm 1.2$, $z = 0.56$, $P = 0.57$; overall effect of clade: $\chi^2 = 80.8$, df = 2, $P < 0.0001$). This difference persisted when 'suspected' cases were included (Extended Data Table 5a).

To investigate which factors best explained variation in killing rates among chimpanzee communities, we used an information theoretic approach[22], controlling for years of observation. We considered three variables for the human impact hypothesis: provisioned (whether the community had been artificially fed); area (size of protected area, with smaller areas assumed to experience more impacts); and disturbance. We also considered three variables for the adaptive strategies hypothesis: clade (eastern and western chimpanzees may have different histories of selection for violence); males (number of adult males, which may influence rates of killing via intensity of reproductive competition and/or coalitional fighting power), and density (number of individuals per km^2, which may affect frequency of intercommunity encounter and/or intensity of resource competition). We consider density to reflect natural food abundance. For example, at Ngogo (4.5 chimpanzees per km^2), vegetation sampling revealed high forest productivity[23] and chimpanzees have high C-peptide levels[24], indicating high energy balance; whereas at Fongoli (0.37 chimpanzees per km^2), chimpanzees range widely across a dry savannah with sparse food[25]. Density was unrelated to disturbance (general linear model, $F_{1,16} = 1.4$, $P = 0.26$).

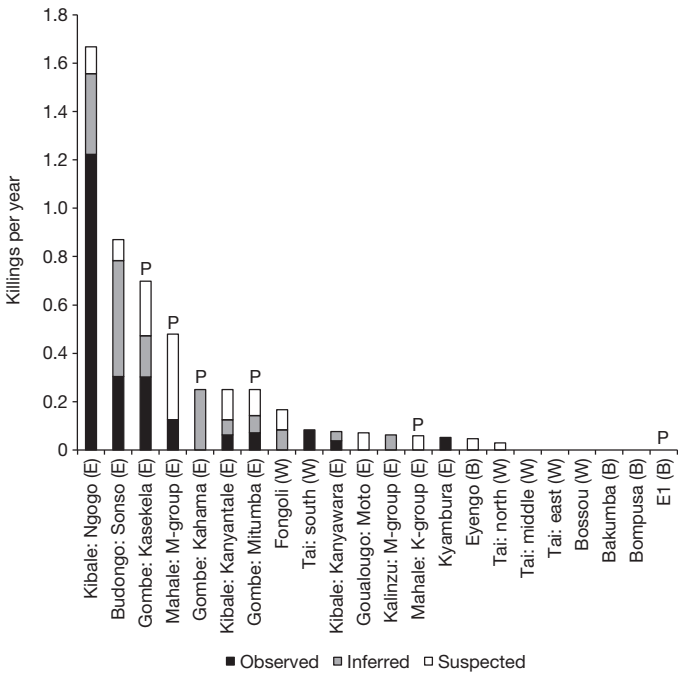


Figure 1 | Number of victims killed per year by members of study communities. Bars indicate the annual rate of observed (black), inferred (grey), and suspected (white) killings by each community for bonobos (B; $n = 4$), eastern chimpanzees (E; $n = 12$), and western chimpanzees (W; $n = 6$). Communities with a history of provisioning are indicated by (P).

Of the 16 models we considered (Table 3), four of the five models in the resulting 95% confidence set included combinations of the adaptive variables; the fifth model included the three human impact variables. The best model included only males and density, and was supported 6.8 times more strongly than the human impact model (evidence ratio = $w_i/w_j = 0.40/0.059 = 6.8$). Considering model-averaged parameter estimates[22], increases in *males* and *density* increased the number of killings; for all other parameter estimates, the 95% confidence intervals included zero (Table 3 and Fig. 2). Excluding one community (Ngogo) that had both an unusually high killing rate and unusually many males resulted in similar values for model-averaged parameters, but only the estimate for density excluded zero from the 95% confidence interval (Extended Data Table 5b; $n = 17$).

Opposite to predictions from the human impact hypothesis (Table 2), provisioned and disturbance both had negative effects; the estimates for these parameters included zero in the 95% confidence intervals (Table 3 and Extended Data Fig. 2b). The highest rate of killing occurred at a relatively undisturbed and never-provisioned site (Ngogo); chimpanzees at the least disturbed site (Goualougo) were suspected of one killing and inferred to have suffered an intercommunity killing; and no killings occurred at the site most intensely modified by humans (Bossou).

As a test of confidence, we investigated the effects of including suspected cases and data from bonobos. Including suspected cases changed western and provisioned from negative to positive (Extended Data Table 5b). Nonetheless, even with these suspected cases, none of the estimates for human impact variables excludes zero from the 95% confidence interval. Including bonobo data widened the confidence intervals for density (Extended Data Table 5b), probably because two bonobo communities had high densities (Extended Data Fig. 1a). With either suspected cases or bonobo data added, only for males did the 95% confidence intervals exclude zero (Extended Data Table 5b). Thus, although demographic variables explain variation in rates of killing better than human impact variables, the confidence intervals are sensitive to including suspected cases or data from another species (bonobos).

These analyses combine killings committed for varied reasons by individuals in different age–sex classes. A full explanation of these events requires a finer grained analysis. To this end, we examined variation over time and among different categories of attacker and victim.

Increasing human impacts have been proposed to cause increasing numbers of killings in recent years[8]. However, controlling for changes in the number of communities observed per year (communities), the rate of killing has not changed over time (year). Using an information theoretic approach[22] to compare three different models (year; communities; and year plus communities), the best model contained only communities; considering model-averaged parameters, the 95% confidence interval excluded zero for communities, but not year (Poisson regression: $n = 52$ years; model-averaged parameters and 95% confidence interval: $\beta_0 = 10$ (−38 to 58); $\beta_{year} = -0.0058$ (−0.022 to 0.010); $\beta_{communities} = 0.18$ (0.10–0.26); Extended Data Table 5c).

Table 3 | Summary of model selection: number of killings per community

	B	Clade	Males	Density	Area	Prov	Dist	*K*	Δ_i	w_i
	−3.6		0.081	0.21				4	0.00	0.40
	−2.3	−1.9	0.073					4	0.61	0.30
	−3.1	−1.4	0.073	0.15				5	1.8	0.16
	−2.7		0.087					3	3.4	0.07
	7.1				−0.0016	−1.4	−0.63	5	3.8	0.06
	−2.2	2.4	0.10	0.42	−0.00083	1.3	−0.27	8	10	0.00
	3.7				−0.0011		−0.40	4	12	0.00
	−2.0	−2.1		0.17				4	17	0.00
	−1.2	−2.7						3	18	0.00
	−2.8			0.28				3	21	0.00
	−1.1				−0.00042			3	24	0.00
	−1.1				−0.00042	−0.12		4	28	0.00
	−1.5							2	34	0.00
	−1.6					0.19		3	36	0.00
	−1.4						−0.011	3	37	0.00
	−1.6					0.18	−0.0046	4	40	0.00
MAP	−2.4	−0.78	0.073	0.11	−0.00010	−0.078	−0.038			
2.5%	−5.0	−1.8	0.053	0.00029	−0.00027	−0.24	−0.11			
97.5%	0.12	0.25	0.093	0.22	0.000083	0.082	0.033			

Parameters include the intercept (*B*); impact of western relative to the eastern clade of chimpanzees; mean number of adult males per community (males); mean population density per community (density); size of protected area in km^2 (area); history of regular provisioning with food (prov); disturbance rating (dist.); the number of free parameters (*k*) including the dispersion parameter ($\hat{c} = 2.8$); the difference in Akaike information criterion (corrected for overdispersion: QAICc) between the *i*th model and the best model (Δ_i); and model weight (w_i). Models are arranged in order from best (lowest ΔQAICc$_i$) to worst (highest ΔQAICc). The weight of the model (w_i) is the probability that a given model is the best model in a given set of models. Model-averaged parameter estimates (MAP) with upper (97.5%) and lower (2.5%) bounds of the 95% confidence intervals are given in the bottom rows.

Killings involved a median of five male attackers (range: 0–19) and no females (range: 0–6). Considering all cases for which the number of attackers was observed ($n = 58$) or could be inferred ($n = 6$), males constituted 92% of participants in attacks (338/366). Controlling for observation time and community composition, males were much more likely to participate in killings than females (negative binomial mixed model: $n = 36$ observations (fixed effects: *sex* with 2 levels; random effects: community with 18 levels); $\beta_0 = -6.9 \pm 0.98$; $\beta_{\text{males}} = 2.6 \pm 0.59$, $z = 4.42$, $P < 0.0001$). Females sometimes joined males in attacking grown individuals ($n = 3$), but when acting without males, females killed only young infants ($n = 8$).

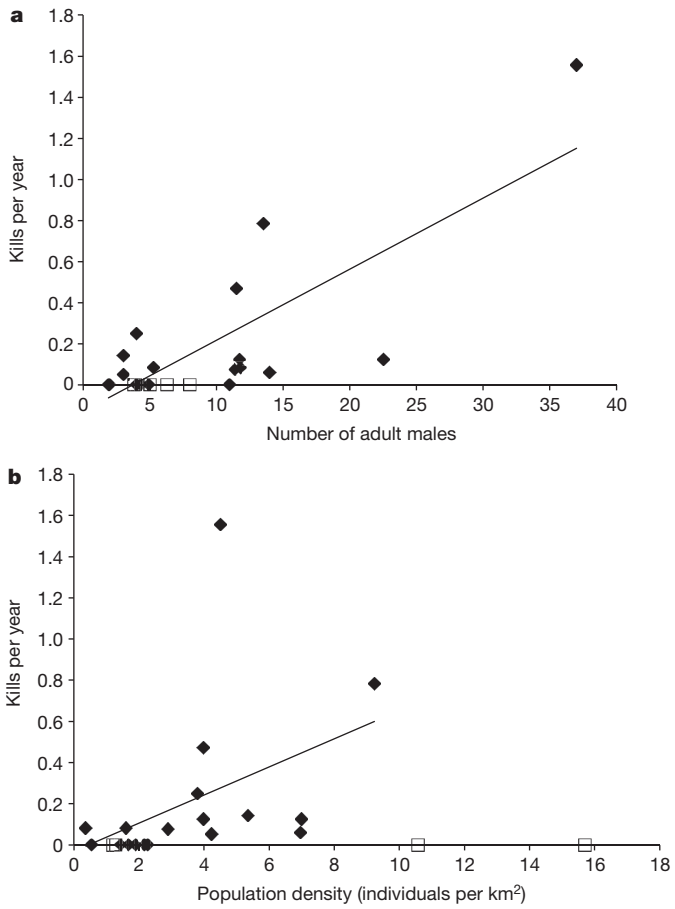


Figure 2 | Number of killings per year for each community versus number of males and population density. Rates for each community are indicated by black diamonds (chimpanzees; $n = 18$) and open squares (bonobos; $n = 4$). Black lines indicate simple linear regression for chimpanzee data for illustrative purposes only; statistical tests were done using Poisson regressions. **a**, Number of killings versus number of males. **b**, Number of killings versus population density (individuals per km^2).

Controlling for observation time and community composition, males and infants had the highest probability of being killed (Extended Data Table 6). Notably, during infanticides, attackers sometimes removed infants from mothers under circumstances in which they appeared capable of killing the mother as well, but did not do so.

Most victims were members of different communities from the attackers ($n = 62$ of 99 cases; 63%) and thus not likely to be close kin[26]. This difference is particularly striking given that chimpanzees could potentially attack members of their own community on a daily basis, but rarely encounter members of other communities (for example, 1.9% of follow days at Kanyawara[27]).

Intercommunity killings mainly involved parties with many males (median = 9 males, range: 2–28, $n = 36$ cases with known numbers of attackers) attacking isolated or greatly outnumbered males or, more often, mothers with infants (median = 0 males, range: 0–3, $n = 30$; median = 1 female, range: 0–5, $n = 31$). For 30 cases in which the number of adult and adolescent males and females on each side were known, attackers outnumbered defenders by a median factor of 8 (range: 1–32; Extended Data Table 7). Most intercommunity killings thus occurred when attackers overwhelmingly outnumbered victims.

Several robust patterns emerge from these data. Killing was most common in eastern chimpanzees and least common among bonobos. Among chimpanzees, killings increased with more males and higher population density, whereas none of the three human impact variables had an obvious effect. Male chimpanzees killed more often than females, and killed mainly male victims; attackers most frequently killed unweaned infants; victims were mainly members of other communities (and thus unlikely to be close kin); and intercommunity killings typically occurred when attackers had an overwhelming numerical advantage. The most important predictors of violence were thus variables related to adaptive strategies: species; age–sex class of attackers and victims; community membership; numerical asymmetries; and demography. We conclude that patterns of lethal aggression in *Pan* show little correlation with human impacts, but are instead better explained by the adaptive hypothesis that killing is a means to eliminate rivals when the costs of killing are low.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 3 March; accepted 4 August 2014.

1. Goodall, J. *The Chimpanzees of Gombe: Patterns of Behavior* (Belknap Press, 1986).
2. Wrangham, R. W. Evolution of coalitionary killing. *Am. J. Phys. Anthropol.* **110,** (suppl.) 1–30 (1999).
3. Wilson, M. L. & Wrangham, R. W. Intergroup relations in chimpanzees. *Annu. Rev. Anthropol.* **32,** 363–392 (2003).
4. Boesch, C. *The Real Chimpanzee: Sex Strategies in the Forest* (Cambridge Univ. Press, 2009).
5. Mitani, J. C., Watts, D. P. & Amsler, S. J. Lethal intergroup aggression leads to territorial expansion in wild chimpanzees. *Curr. Biol.* **20,** R507–R508 (2010).
6. Power, M. *The Egalitarians—Human and Chimpanzee: An Anthropological View of Social Organization* (Cambridge Univ. Press, 1991).
7. Sussman, R. W. in *War, Peace, and Human Nature: The Convergence of Evolutionary and Cultural Views* (ed. Fry, D. P.) 97–111 (Oxford Univ. Press, 2013).
8. Ferguson, R. B. in *Origins of Altruism and Cooperation* (eds Sussman, R. W. & Cloninger, C. R.) 249–270 (2011).
9. Bartlett, T. Q., Sussman, R. W. & Cheverud, J. M. Infant killing in primates: a review of observed cases with specific reference to the sexual selection hypothesis. *Am. Anthropol.* **95,** 958–990 (1993).
10. Mitani, J. C. Cooperation and competition in chimpanzees: current understanding and future challenges. *Evol. Anthropol.* **18,** 215–227 (2009).
11. Boesch, C. *et al.* Intergroup conflicts among chimpanzees in Tai National Park: lethal violence and the female perspective. *Am. J. Primatol.* **70,** 519–532 (2008).
12. Wrangham, R. W., Wilson, M. L. & Muller, M. N. Comparative rates of violence in chimpanzees and humans. *Primates* **47,** 14–26 (2006).
13. Williams, J. M., Oehlert, G., Carlis, J. & Pusey, A. E. Why do male chimpanzees defend a group range? Reassessing male territoriality. *Anim. Behav.* **68,** 523–532 (2004).
14. Mitani, J. C. Demographic influences on the behavior of chimpanzees. *Primates* **47,** 6–13 (2006).
15. Fawcett, K. & Muhumuza, G. Death of a wild chimpanzee community member: possible outcome of intense sexual competition. *Am. J. Primatol.* **51,** 243–247 (2000).
16. Watts, D. P. Intracommunity coalitionary killing of an adult male chimpanzee at Ngogo, Kibale National Park, Uganda. *Int. J. Primatol.* **25,** 507–521 (2004).
17. Pusey, A. E. *et al.* Severe aggression among female *Pan troglodytes schweinfurthii* at Gombe National Park, Tanzania. *Int. J. Primatol.* **29,** 949–973 (2008).
18. Hamilton, W. D. The genetical evolution of social behavior. I, II. *J. Theor. Biol.* **7,** 1–16 (1964).
19. Maynard Smith, J. The theory of games and the evolution of animal conflicts. *J. Theor. Biol.* **47,** 209–221 (1974).
20. Prado-Martinez, J. *et al.* Great ape genetic diversity and population history. *Nature* **499,** 471–475 (2013).
21. Boesch, C., Hohmann, G. & Marchant, L. F. *Behavioral Diversity in Chimpanzees and Bonobos* (Cambridge Univ. Press, 2002).
22. Burnham, K. P. & Anderson, D. R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd edn, xxvi, 488 (Springer, 2002).
23. Potts, K. B., Watts, D. P. & Wrangham, R. W. Comparative feeding ecology of two communities of chimpanzees (*Pan troglodytes*) in Kibale National Park, Uganda. *Int. J. Primatol.* **32,** 669–690 (2011).
24. Emery Thompson, M., Muller, M. N., Wrangham, R. W., Lwanga, J. S. & Potts, K. B. Urinary C-peptide tracks seasonal and individual variation in energy balance in wild chimpanzees. *Horm. Behav.* **55,** 299–305 (2009).
25. Sponheimer, M. *et al.* Do "savanna" chimpanzees consume C-4 resources? *J. Hum. Evol.* **51,** 128–133 (2006).
26. Inoue, E., Inoue-Murayama, M., Vigilant, L., Takenaka, O. & Nishida, T. Relatedness in wild chimpanzees: influence of paternity, male philopatry, and demographic factors. *Am. J. Phys. Anthropol.* **137,** 256–262 (2008).
27. Wilson, M. L., Kahlenberg, S. M., Wells, M. T. & Wrangham, R. W. Ecological and social factors affect the occurrence and outcomes of intergroup encounters in chimpanzees. *Anim. Behav.* **83,** 277–291 (2012).

Acknowledgements This study was funded by National Science Foundation grants BCS-0648481 and LTREB-1052693 and National Institutes of Health grant R01 AI 058715. Numerous additional sources of funding have supported the long-term studies that contributed data to this study. We thank J. H. Jones for statistical advice; L. Pintea for preparing the map for Extended Data Fig. 1b; I. Lipende and R. Lawrence for providing details on recent cases at Gombe and Kanyantale; S. Amsler for helping to calculate the range of the Kanyantale community, and the many field assistants who collected data.

Author Contributions All authors contributed to the acquisition, analysis and interpretation of the data; M.L.W., R.W.W., and J.C.M. initiated and conceived the study; M.L.W. and R.M. performed statistical analyses; C.B., B.F., T.F., C.H., C.L.H., G.H., N.I., K.K., J.N.L., T.M., J.C.M., D.C.M., D.M., M.N.M., M.N., J.P., A.E.P., C.S., N.S., D.P.W., F.W., K.Z., M.L.W., R.M.W., and R.W.W. conducted and supervised fieldwork; C.B., T.F., I.C.G., C.H., C.L.H., G.H., J.N.L., T.M., J.C.M., D.C.M., D.M., M.N.M., M.N., J.P., J.R., C.S., A.M.S., N.S., M.L.W., M.W., D.P.W., F.W., R.W.W. and K.Z. provided demographic and ranging data; C.B., T.F., C.H., G.H., J.N.L., T.M., J.C.M., M.N., J.P., A.E.P., N.S., F.W., M.L.W., R.W.W., and K.Z. provided data on site characteristics and human disturbance ratings; M.L.W. coordinated the contributions of all authors; M.L.W. wrote the paper with J.C.M., D.P.W., R.W.W. and input from all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.L.W. (wilso198@umn.edu).

LETTER

doi:10.1038/nature13469

Optimization of lag time underlies antibiotic tolerance in evolved bacterial populations

Ofer Fridman[1], Amir Goldberg[1], Irine Ronin[1], Noam Shoresh[2] & Nathalie Q. Balaban[1]

The great therapeutic achievements of antibiotics have been dramatically undercut by the evolution of bacterial strategies that overcome antibiotic stress[1,2]. These strategies fall into two classes. 'Resistance' makes it possible for a microorganism to grow in the constant presence of the antibiotic, provided that the concentration of the antibiotic is not too high. 'Tolerance' allows a microorganism to survive antibiotic treatment, even at high antibiotic concentrations, as long as the duration of the treatment is limited. Although both resistance and tolerance are important reasons for the failure of antibiotic treatments[3–6], the evolution of resistance[7–9] is much better understood than that of tolerance. Here we followed the evolution of bacterial populations under intermittent exposure to the high concentrations of antibiotics used in the clinic and characterized the evolved strains in terms of both resistance and tolerance. We found that all strains adapted by specific genetic mutations, which became fixed in the evolved populations. By monitoring the phenotypic changes at the population and single-cell levels, we found that the first adaptive change to antibiotic stress was the development of tolerance through a major adjustment in the single-cell lag-time distribution, without a change in resistance. Strikingly, we found that the lag time of bacteria before regrowth was optimized to match the duration of the antibiotic-exposure interval. Whole genome sequencing of the evolved strains and restoration of the wild-type alleles allowed us to identify target genes involved in this antibiotic-driven phenotype: 'tolerance by lag' (*tbl*). Better understanding of lag-time evolution as a key determinant of the survival of bacterial populations under high antibiotic concentrations could lead to new approaches to impeding the evolution of antibiotic resistance.

We exposed batch cultures of the bacterium *Escherichia coli* to a high concentration of ampicillin (100 μg ml^{-1}), which is 15 times greater than the minimum inhibitory concentration (MIC), in a cyclic manner. The frequency of mutants that were resistant to this concentration[10] was less than 10^{-11}. In each cycle (Fig. 1a), parallel overnight cultures in small volumes were resuspended in fresh medium containing ampicillin. After a fixed duration, T_a, of daily exposure to the antibiotic, the cultures were washed to remove the drug, resuspended in fresh medium and grown overnight. We evolved six populations, two for each of the three exposure durations: $T_a = 3$, 5 and 8 h. In all cases, the bacteria soon adapted to the stressful regimen, and follow-up experiments established the genetic basis of this adaptation.

In the first cycle, the proportion of surviving bacteria at each T_a was less than 0.1%. After eight to ten cycles, ampicillin became at least an order of magnitude less effective at killing the bacteria (Fig. 1b and Extended Data Fig. 1). The increased survival rate of the evolved bacteria could have been achieved by a mutation that conferred resistance. We found, however, that the MIC of ampicillin for clones isolated from the evolved lines (*tbl3a*, *tbl5a* and *tbl8a*, evolved at $T_a = 3$, 5 and 8 h, respectively) was indistinguishable from that for their ancestors (Fig. 1c). Similar to the MIC, the measure MDK_{99} (the minimum duration for killing 99% of cells) may be defined to quantify tolerance (Box 1). A higher tolerance translates to a longer MDK_{99}: that is, a treatment of longer duration is needed to reach the same level of killing. We measured the MDK_{99} for the wild-type and evolved strains and found that the MDK_{99} of the evolved population increased with the duration of the stress period, reaching values as high as 15 times the MDK_{99} of the wild-type strain (Fig. 1d). We conclude that these populations have all adapted to the antibiotic regimen through tolerance and not resistance.

Different mechanisms can enable bacteria to endure an interval of exposure to antibiotics. One way to achieve tolerance is by slowed growth[11]. However, no reduction in growth that could account for the observed tolerance was measured (Extended Data Fig. 2a). Another strategy for tolerance is related to the cells being in stationary phase before exposure to antibiotics, resulting in a delay in regrowth when switched to a new environment. By extending the time to first division (the single-cell

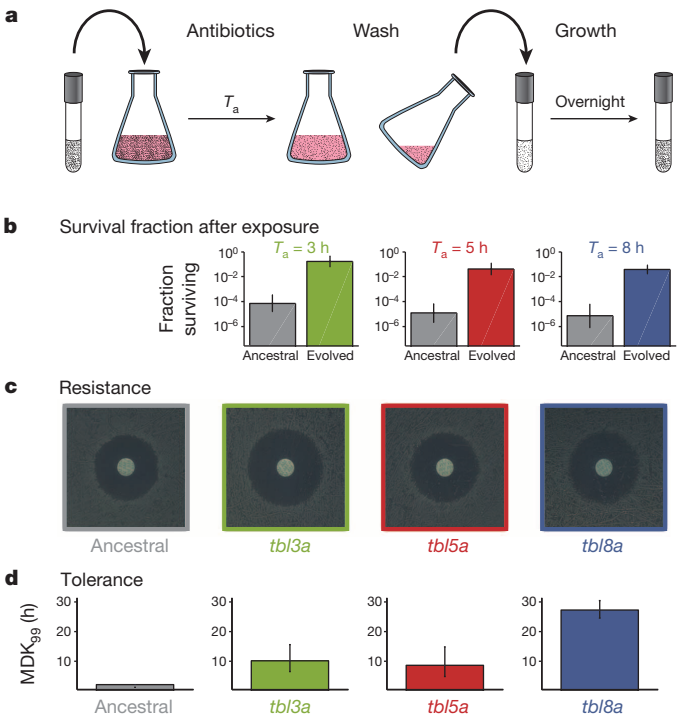


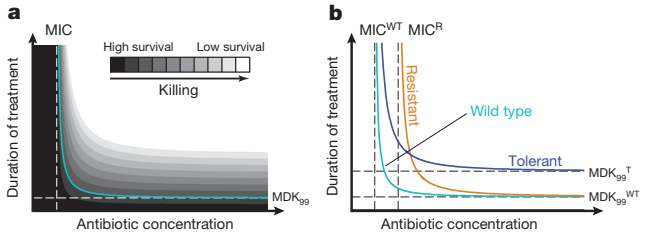
Figure 1 | Resistance and tolerance of the evolved strains. **a**, Experimental design for cyclic exposure to antibiotics. In each cycle, a small-volume overnight culture was resuspended in a larger volume of fresh medium containing 100 μg ml^{-1} ampicillin for an exposure time T_a. After the antibiotic was washed out, the culture was resuspended in fresh medium and grown overnight. **b**, Survival of the evolved strains and ancestral strain after 100 μg ml^{-1} ampicillin treatment for $T_a = 3$, 5 or 8 h: strains *tbl3a*, *tbl5a* and *tbl8a*, respectively (see Extended Data Table 1). Data are presented as the mean ± s.d. of two independent experiments. **c**, MIC test carried out using disc diffusion antibiotic sensitivity testing. **d**, Increase in tolerance of the evolved strains. MDK_{99} (see Box 1) was determined by measuring the time to kill 99% of the population. Data are presented as the mean ± s.d. from four experiments.

[1]Racah Institute of Physics, The Sudarsky Center for Computational Biology and the Center for NanoScience, Edmond J. Safra Campus, The Hebrew University, Jerusalem 91904, Israel. [2]Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA.

BOX 1

Schematics of resistance and tolerance to bactericidal antibiotics

An antibiotic treatment is characterized by the concentration at which it is administered and its duration. The outcome of a treatment can be quantified by the fraction of bacteria killed by the drug. The curves (Box 1 Figure, panel a) indicate lines of equal killing for the wild-type (WT) strain. The cyan curve represents the combinations of treatment durations and concentrations that result in 99% killing. The curve displays two asymptotic behaviours: the vertical asymptote shows the concentration below which the culture will not be killed, even for a very long antibiotic treatment, namely the MIC. The horizontal asymptote shows the characteristic time needed to kill 99% of the culture, in the limit of high antibiotic concentrations. We term this asymptotic value the minimum duration for killing, or MDK_{99}. Schematics of the 99% killing curves for WT (cyan), resistant (R, orange) and tolerant (T, blue) strains are shown in Box 1 Figure, panel b. An increase in resistance translates into a shift in the MIC to higher concentrations, whereas an increase in tolerance manifests as a shift in the MDK_{99} to longer treatment durations.



lag time[12]) and remaining longer in a dormant state, a cell may avoid the harms of antibiotics. A case in point is the tolerance of type I persistent bacteria[2], which is based on the existence of a subpopulation of cells with sufficiently long single-cell lag times[13]. To explore the possibility of the evolution of a lag-time-related strategy, we directly monitored, under the microscope, the time to first division of single cells taken from an overnight culture in stationary phase (Fig. 2a). Whereas cells from the ancestral clone divided within half an hour, the lag times of the evolved cells were distributed over many hours.

To corroborate that extended lag time is the main factor reducing the killing rate by ampicillin in our protocol, we measured the survival of evolved clones under exposure to norfloxacin, a drug that belongs to a different class of bactericidal antibiotic. The expectation was that the benefits of the slow exit from the dormant phase are generic and would carry over to a different drug that targets growing bacteria[14,15]. Indeed, a comparable increase in tolerance to norfloxacin was observed for the evolved clones, ruling out the possibility that a specific ampicillin-associated mechanism was responsible for the increase in tolerance (Extended Data Fig. 2b). Furthermore, the protective effect of the lag time was negated when the evolved clones were maintained in the exponential growth phase, a condition under which neither stationary phase nor the subsequent lag phase are reached (Fig. 2b). These results establish that in our cyclic protocol, the adaptive trait conferring tolerance to antibiotic treatments is the extended single-cell lag time. We term this adaptation 'tolerance by lag'(*tbl*), to distinguish it from tolerance that may be due to other factors.

To quantify the extension of single-cell lag times at the population level, we measured the distribution of these times with our recently developed ScanLag set-up[16]. Measuring the lag times of hundreds of cells in each case, we used this method to obtain single-cell lag-time distributions for the ancestral population and for clones isolated from the end populations evolved under the three cyclic regimens (Fig. 2c, Supplementary Video 1 and Extended Data Fig. 3). For each of the empirical distributions,

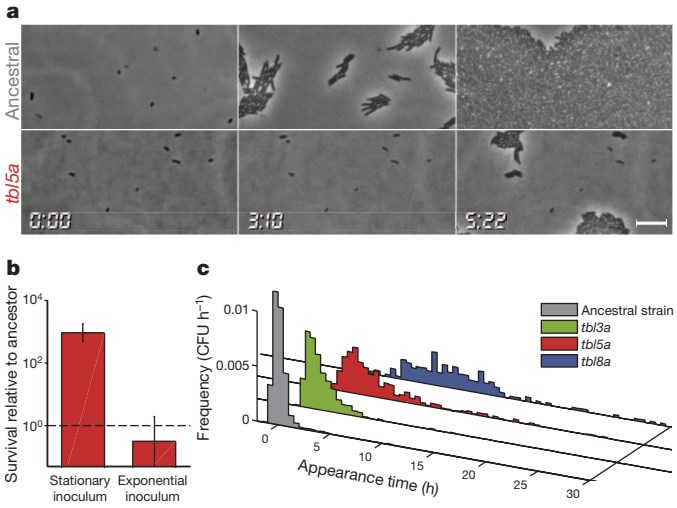


Figure 2 | Changes in the single-cell lag-time distributions underlie antibiotic tolerance. **a**, Phase-contrast images of time-lapse microscopy of single bacteria from a stationary culture plated on fresh medium show the extended lag of the evolved strain. Times are indicated in hours and minutes. Scale bar, 5 μm. **b**, Survival advantage of the evolved strain (*tbl5a*) over the ancestral strain (shown as fold change) when exposed to ampicillin, after stationary phase or after exponential growth. The dashed line indicates ancestral survival. Data are presented as the mean ± s.d. of three independent experiments. **c**, Measurement of the lag-time distribution with ScanLag. The appearance time of colonies was continuously monitored by an automated scanner system[16]. The histograms show the proportion of colony-forming units (CFUs) detected at each time point (mean = 0.78 h (ancestral), 2.1 h (*tbl3a*), 4.9 h (*tbl5a*), 10.3 h (*tbl8a*); median = 0.7 h (ancestral), 1.7 h (*tbl3a*), 3.7 h (*tbl5a*), 9.2 h (*tbl8a*), with s.e.m. below 8% and sample sizes of $n = 514, 747, 423$ and 168, respectively). Colonies appearing at later times grow at the same rate (Extended Data Fig. 2a). Stationary cultures were grown from a single colony isolated from the majority clone of the end populations.

we calculated the mean single-cell lag time, τ_{lag}. We found that τ_{lag} increased from 1.0 ± 0.2 h for the ancestral strain to several hours for the evolved strains. Specifically, τ_{lag} was found to be 3.4 ± 1.0 h, 5.1 ± 0.2 h and 10 ± 1 h for the clones that evolved in response to cyclic 3, 5 and 8 h exposures to antibiotic, respectively. These data show that the bacterial cultures under cyclic exposure to antibiotics adapted by extending the typical timescale of exit from the lag phase to approximately match the timescale of the antibiotic exposure.

While delaying the exit from lag phase provides protection from the antibiotic, a delay that is too long comes at the expense of time that could be spent proliferating once the antibiotics have been washed out. We explored this trade-off computationally, using a simple population-dynamics model that describes the growth and death of cells under cyclic antibiotic exposure (Fig. 3a–c and see Methods). By approximating the empirical single-cell lag-time distribution as exponential, the fitness of a population is a function of two parameters: the mean lag time, τ_{lag}, and the antibiotic-exposure time, T_a. We used the model to calculate the value of τ_{lag} that maximizes the fitness for a fixed T_a, and we found that the optimal lag time increases linearly with T_a, which is in good agreement with the empirical results (Fig. 3d).

We examined the genetic basis of the *tbl* phenotype by sequencing clones derived from the evolved populations (see Supplementary Methods)[17]. In each clone, few mutations were found, with a total of eight mutations across all sequenced clones being confirmed by Sanger sequencing. All mutations were in coding regions and were non-synonymous, suggesting that the affected proteins have a functional role in the *tbl* phenotype. To explore this possibility, we restored the wild-type alleles in the evolved strains and identified which genes restored the wild-type lag time; we defined these genes as *tbl* genes (Extended Data Fig. 4). Of the six genes affected by the eight mutations, three were found to be *tbl* genes by this definition (Extended Data Tables 1 and 2). Notably, we observed

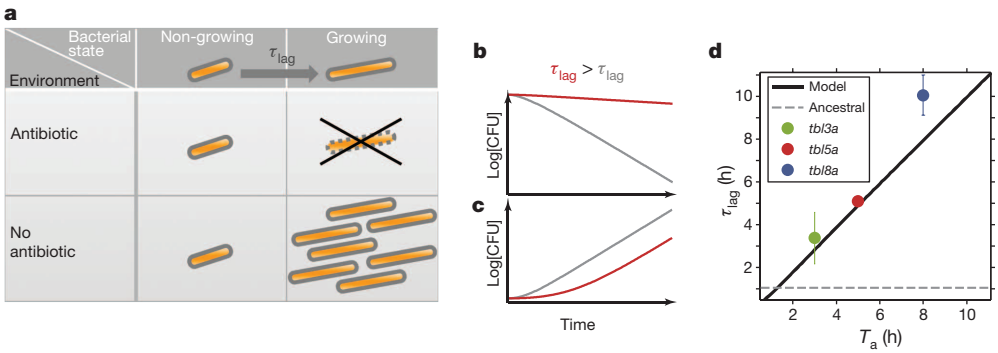


Figure 3 | Optimization of lag time. **a**, Schematic illustration of the processes in the theoretical model for the optimization of lag time. The fate of bacteria is determined by their growth state and by the environment. The non-growing phase (lag) protects the bacteria from the antibiotics. **b**, **c**, Non-growing bacteria switch to the growing state with a typical timescale of τ_{lag}. The trade-off of extending the lag time is illustrated by two strains: one with a short τ_{lag} (grey), and the other with a long τ_{lag} (red). In the presence of antibiotics, the longer τ_{lag} reduces the killing and increases tolerance (**b**). In the absence of antibiotics, the longer τ_{lag} delays the resumption of growth and bears a fitness cost (**c**). **d**, Lag time as a function of duration of antibiotic treatment, T_a. The empirical mean lag-time dependence (circles) of the evolved *tbl* clones follows the model predictions for lag-time optimization (solid line). Data are presented as the mean ± s.d. of three independent experiments. The dashed line is the mean lag time of the ancestral strain.

different mutations within the same gene in different clones, indicating that these genes are under strong selection. All *tbl* genes were found to have fixed mutations in the evolved populations (Extended Data Table 3).

The *tbl* gene candidates are associated with several pathways. Two of these pathways, toxin–antitoxin modules[18,19] and aminoacyl-transfer RNA synthetase[20–22], have been implicated in increased persistence. Interestingly, quantitative analysis has shown that toxin–antitoxin modules are a network motif that is ideally suited to set the timescale of the single-cell lag-time distribution[23]. In contrast to these pathways, no relationship between the third emerging pathway, *prs*, and either lag or tolerance has been described in the literature. More experiments are needed to understand the relationship between the essential *prs* gene and the *tbl* phenotype.

The evolution of the lag-time distribution is a remarkable demonstration of an adaptation not to the specific nature of an environmental stress but to its duration. Not only does this finding illustrate that temporal parameters associated with growth can be readily changed during evolution[24], but also the correspondence between stress duration and the timescale of the evolved phenotypes shows how fine-tuned this adaptation can be. It is also notable that the malleable trait here is the lag time. Lag is commonly viewed as an inevitable delay in adjustment to new conditions imposed by metabolic constraints. Our study shows this phenotype in a different light: extended lag time as an advantageous trait for avoiding growth in adverse conditions and a trait that can evolve to accommodate selective pressures.

In fact, it is the entire distribution of single-cell lag times that sets the fitness of the population, and it is the distribution as a whole that is being shaped during evolution. In a short period, the populations in this study adapted to antibiotic pressure by extending the single-cell lag-time distribution (Fig. 2c). While the shift in the mean lag time is beneficial, this is not the case for the corresponding increase in variance. Indeed, for our regimen, maximum fitness is reached if all cells exit lag together, at the end of the antibiotic-exposure period. The observed increase in both the mean and the variance of the lag-time distribution is interesting and may indicate that those parameters are constrained at the molecular level to evolve concomitantly. Such an increase in variance could itself reflect past selection for a bet-hedging strategy[25,26], as variance in the single-cell lag times does make sense when the duration of the antibiotic-exposure interval is not completely predictable. More generally, the timescales and stochasticity of the environmental sequence of conditions[27] may shape the single-cell lag-time distribution beyond changes in the mean and variance[26], as in the case of type I persistence to antibiotic treatments, in which only a fraction of the population is tolerant through having a long lag time[13]. It should be noted that in our evolved strains, the whole population becomes tolerant (Extended Data Fig. 1).

From a clinical point of view, tolerance by lag presents a major challenge[28]: in contrast to the specificity of resistance, the *tbl* phenotype confers a survival advantage against a broad spectrum of drugs and stresses. The revised understanding of antimicrobial tolerance offered here suggests new ways to eliminate bacterial populations, such as interfering with the precursor signal (stationary phase) from which bacteria anticipate stress. In addition, it is possible that tolerance facilitates the subsequent evolution of antibiotic resistance and that reducing tolerance might impede the emergence of antibiotic resistance. Understanding tolerance, resistance and the interplay between them in the adaptation of microorganisms to drugs is a critical goal in addressing the decreased efficacy of antibiotics[29].

METHODS SUMMARY

Evolutionary protocol of cyclic exposure to antibiotics. Evolution under cyclic antibiotic exposure consisted of three steps. First, an overnight culture (0.5 ml; 1×10^9 bacteria) was resuspended in 50 ml LB broth Lennox (LBL) supplemented with 100 μg ml^{-1} ampicillin (Sigma) and incubated at 37 °C with shaking at 300 r.p.m. for T_a hours ($T_a = 3, 5$ or 8 h). Second, the antibiotic-containing medium was removed by washing twice in LBL (10 min centrifugation at 1,400*g*). Last, the culture was resuspended in 1 ml fresh LBL and grown overnight (for approximately 20, 18 and 15 h, for the $T_a = 3, 5$ and 8 h cultures, respectively) at 37 °C with shaking. It should be noted that nearly all of the surviving cells were kept from cycle to cycle, thus minimizing drift; this is in contrast to typical serial dilution evolution experiments, in which most of the culture is discarded[30].

Single-cell lag-time distribution measurement with ScanLag. Overnight cultures grown from a single colony were plated at appropriate dilutions on solid LBL agar medium supplemented with 12.5 μg ml^{-1} chloramphenicol (Sigma). A custom, automated scanner array set-up was used to monitor the appearance of thousands of colonies on plates over several days and to extract lag-time distributions by automated image analysis[16]. Single-cell microscopic observations confirmed that the delay in colony appearance was due to an extended lag.

Theoretical model. The model describes a bacterial population consisting of two phenotypes, lagging (L) and growing (G) bacteria, whose dynamics are governed by first-order, linear differential equations: $\dot{G} = \alpha\,(G/\tau_{grow}) + (L/\tau_{lag})$ and $\dot{L} = -(L/\tau_{lag})$, where $1/\tau_{lag}$ is the rate at which lagging bacteria switch to growth and $1/\tau_{grow}$ is the growth rate. The outcome depends on the environment and is captured by the parameter α: in the absence of antibiotics, division leads to population growth and $\alpha = 1$; however, when antibiotics are present, the cells that try to divide die, and $\alpha = -1$.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 1 November 2013; accepted 12 May 2014.
Published online 25 June 2014.

1. Bush, K. *et al.* Tackling antibiotic resistance. *Nature Rev. Microbiol.* **9,** 894–896 (2011).

2. Balaban, N., Merrin, J., Chait, R., Kowalik, L. & Leibler, S. Bacterial persistence as a phenotypic switch. *Science* **305,** 1622–1625 (2004).
3. Collignon, P. Antibiotic resistance. *Med. J. Aust.* **177,** 325–329 (2002).
4. Spellberg, B., Powers, J., Brass, E., Miller, L. & Edwards, J. Trends in antimicrobial drug development: implications for the future. *Clin. Infect. Dis.* **38,** 1279–1286 (2004).
5. Fauvart, M., De Groote, V. & Michiels, J. Role of persister cells in chronic infections: clinical relevance and perspectives on anti-persister therapies. *J. Med. Microbiol.* **60,** 699–709 (2011).
6. Lewis, K. Persister cells. *Annu. Rev. Microbiol.* **64,** 357–372 (2010).
7. Toprak, E. *et al.* Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. *Nature Genet.* **44,** 101–105 (2012).
8. Gullberg, E. *et al.* Selection of resistant bacteria at very low antibiotic concentrations. *PLoS Pathogens* **7** (2011).
9. Lee, H., Molla, M., Cantor, C. & Collins, J. Bacterial charity work leads to population-wide resistance. *Nature* **467,** 82–85 (2010).
10. Olsson, O., Bergström, S., Lindberg, F. & Normark, S. *ampC* β-lactamase hyperproduction in *Escherichia coli*: natural ampicillin resistance generated by horizontal chromosomal DNA transfer from *Shigella*. *Proc. Natl Acad. Sci. USA* **80,** 7556–7560 (1983).
11. Gilbert, P., Collier, P. & Brown, M. Influence of growth rate on susceptibility to antimicrobial agents: biofilms, cell cycle, dormancy, and stringent response. *Antimicrob. Agents Chemother.* **34,** 1865–1868 (1990).
12. Baranyi, J. Stochastic modelling of bacterial lag phase. *Int. J. Food Microbiol.* **73,** 203–206 (2002).
13. Gefen, O. & Balaban, N. The importance of being persistent: heterogeneity of bacterial populations under antibiotic stress. *FEMS Microbiol. Rev.* **33,** 704–717 (2009).
14. Tuomanen, E., Cozens, R., Tosch, W., Zak, O. & Tomasz, A. The rate of killing of *Escherichia coli* by β-lactam antibiotics is strictly proportional to the rate of bacterial growth. *J. Gen. Microbiol.* **132,** 1297–1304 (1986).
15. Wolfson, J., Hooper, D., McHugh, G., Bozza, M. & Swartz, M. Mutants of *Escherichia coli* K-12 exhibiting reduced killing by both quinolone and β-lactam antimicrobial agents. *Antimicrob. Agents Chemother.* **34,** 1938–1943 (1990).
16. Levin-Reisman, I. *et al.* Automated imaging with ScanLag reveals previously undetectable bacterial growth phenotypes. *Nature Methods* **7,** 737–739 (2010).
17. Goecks, J., Nekrutenko, A., Taylor, J. & Galaxy, T. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11,** R86 (2010).
18. Black, D., Kelly, A., Mardis, M. & Moyed, H. Structure and organization of *hip*, an operon that affects lethality due to inhibition of peptidoglycan or DNA synthesis. *J. Bacteriol.* **173,** 5732–5739 (1991).
19. Gerdes, K. & Maisonneuve, E. Bacterial persistence and toxin–antitoxin loci. *Annu. Rev. Microbiol.* **66,** 103–123 (2012).
20. Girgis, H., Harris, K. & Tavazoie, S. Large mutational target size for rapid emergence of bacterial persistence. *Proc. Natl Acad. Sci. USA* **109,** 12740–12745 (2012).
21. Kaspy, I. *et al.* HipA-mediated antibiotic persistence via phosphorylation of the glutamyl-tRNA-synthetase. *Nature Commun.* **4,** 3001 (2013).
22. Germain, E., Castro-Roa, D., Zenkin, N. & Gerdes, K. Molecular mechanism of bacterial persistence by HipA. *Mol. Cell* **52,** 248–254 (2013).
23. Rotem, E. *et al.* Regulation of phenotypic variability by a threshold-based mechanism underlies bacterial persistence. *Proc. Natl Acad. Sci. USA* **107,** 12541–12546 (2010).
24. Oxman, E., Alon, U. & Dekel, E. Defined order of evolutionary adaptations: experimental evidence. *Evolution* **62,** 1547–1554 (2008).
25. Beaumont, H., Gallie, J., Kost, C., Ferguson, G. & Rainey, P. Experimental evolution of bet hedging. *Nature* **462,** 90–93 (2009).
26. Kussell, E., Kishony, R., Balaban, N. & Leibler, S. Bacterial persistence: a model of survival in changing environments. *Genetics* **169,** 1807–1814 (2005).
27. Mitchell, A. *et al.* Adaptive prediction of environmental changes by microorganisms. *Nature* **460,** 220–224 (2009).
28. Nahata, M., Vashi, V., Swanson, R., Messig, M. & Chung, M. Pharmacokinetics of ampicillin and sulbactam in pediatric patients. *Antimicrob. Agents Chemother.* **43,** 1225–1229 (1999).
29. Cohen, N., Lobritz, M. & Collins, J. Microbial persistence and the road to drug resistance. *Cell Host Microbe* **13,** 632–642 (2013).
30. Barrick, J. *et al.* Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* **461,** 1243–1247 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank I. Levin-Riesman for discussions and technical help with the ScanLag system, S. Belkin for strains, and the National BioResource Project and National Institute of Genetics, Japan, for the Keio collection deletion mutants. The work was supported by the European Research Council (Starting Grant no. 260871) and the Israel Science Foundation (no. 592/10). O.F. acknowledges support from the Levtzion Fellowship.

Author Contributions N.Q.B. and O.F. designed the experiments. O.F. performed the experiments. O.F. and N.Q.B. analysed the data. O.F. and N.S. performed the theoretical analysis. O.F. and A.G. analysed the whole genome sequencing data. I.R. made the genetic reconstructions. N.Q.B. and N.S. wrote the manuscript.

Author Information Whole genome sequence data of KLY (ancestral), *tbl3a tbl5a* and *tbl8a*, as well as reconstructed ancestral genome sequence data, have been deposited in the BioProject database under the accession number PRJNA229104. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to N.Q.B. (nathalieqb@phys.huji.ac.il).

LETTER

doi:10.1038/nature13448

Genome sequencing of normal cells reveals developmental lineages and mutational processes

Sam Behjati[1,2], Meritxell Huch[3]*†, Ruben van Boxtel[3]*, Wouter Karthaus[3]*, David C. Wedge[1], Asif U. Tamuri[4], Iñigo Martincorena[1], Mia Petljak[1], Ludmil B. Alexandrov[1], Gunes Gundem[1], Patrick S. Tarpey[1], Sophie Roerink[1], Joyce Blokker[3], Mark Maddison[1], Laura Mudie[1], Ben Robinson[1], Serena Nik-Zainal[1,5], Peter Campbell[1], Nick Goldman[4], Marc van de Wetering[3], Edwin Cuppen[3], Hans Clevers[3] & Michael R. Stratton[1]

The somatic mutations present in the genome of a cell accumulate over the lifetime of a multicellular organism. These mutations can provide insights into the developmental lineage tree[1], the number of divisions that each cell has undergone and the mutational processes that have been operative[2]. Here we describe whole genomes of clonal lines[3] derived from multiple tissues of healthy mice. Using somatic base substitutions, we reconstructed the early cell divisions of each animal, demonstrating the contributions of embryonic cells to adult tissues. Differences were observed between tissues in the numbers and types of mutations accumulated by each cell, which likely reflect differences in the number of cell divisions they have undergone and varying contributions of different mutational processes. If somatic mutation rates are similar to those in mice, the results indicate that precise insights into development and mutagenesis of normal human cells will be possible.

Somatic mutations in normal cells from multicellular organisms can provide insights into the origins and past experiences of each cell[4]. The developmental lineages of individual cells may be reconstructed using insertion or deletion mutations at short tandem repeats, as has been explored previously[1,5–11] (Supplementary Discussion). Complete reconstruction of the bifurcating cell division tree requires at least one somatic mutation in the two daughter cells arising from each cell division. If mutations are generated during mitotic DNA replication, the number of divisions that a cell has undergone may be reflected in the number of mutations present[6,9]. Finally, the processes of DNA damage and/or repair experienced by different cell types may be manifested in distinct mutational signatures within the catalogues of mutation in each cell[2].

Single-cell whole-genome DNA sequencing has the potential to elucidate these aspects of the biology of normal cells. However, technologies are still under development[12] and are often characterized by incomplete genome coverage, suboptimal mutation sensitivity and substantial error rates for most mutation types. Errors mislead lineage reconstruction, distort mutational signatures and cause mis-estimation of numbers of mitoses undergone.

We therefore derived clonal lines from normal mouse cells by using organoid technology[3,13–17] as an alternative to single-cell analysis. Twenty-five organoid lines were obtained from the stomach, small bowel and large bowel of two mice (mouse 1 aged 116 weeks, and mouse 2 aged 98 weeks) and from the prostate of mouse 2 (Supplementary Table 1). The whole genome was sequenced of each line and the polyclonal tail of each mouse, and extensive validation was performed to obtain high-quality catalogues of somatic mutations. After alignment to the reference mouse genome, sets of somatic base substitution mutations were obtained by comparing each single-cell clone with all other clones from the same mouse, and with the tail (see Methods). Analyses of read count frequencies of the somatic substitutions indicated that they were heterozygous and that very few subclonal mutations, which may have arisen *in vitro*, were captured (Extended Data Fig. 1, Supplementary Discussion).

To reconstruct the early developmental lineage tree of each mouse we searched for mutations present in at least two organoids and absent from at least one, applying computational and manual approaches (Methods and Extended Data Figs 2 and 3). All such putative early-embryonic mutations were further assessed by re-sequencing in every organoid, yielding 35 from the two mice (Supplementary Table 2). The phylogenetic lineage was then developed into a hypothetical tree of early embryonic cell divisions (Methods and Extended Data Fig. 3). Of the 23 cell divisions requiring reconstruction to generate a simple bifurcating tree for all the individual cell clones from both mice, 17 were reconstructable and 6 were not (Fig. 1 and Extended Data Fig. 4). Thus the intrinsic substitution mutation rate per cell division in early mouse embryos was almost sufficient to reconstruct the tree. Mutations defining the first four cell generations were found in the tails of the two mice, confirming that they occurred before formation of the three germ layers during gastrulation because they were present in endoderm (the precursor of tissues from which the clonal organoids were derived) and either mesoderm or ectoderm, or both (which contribute to the tail).

The earliest reconstructed cell division in each tree may represent the first division of the fertilized egg. However, it is possible that earlier cell divisions took place, generating daughter cells that were not ancestors of any of the 25 single cells sampled in the two mice. To assess this possibility we examined the sequencing read counts of mutations of the putative first two daughter cells (generation I, cells 'b' and 'c') in the polyclonal tail samples of each mouse. Although we cannot exclude the possibility that a small proportion was from earlier progenitors, the results are compatible with all cells in the tail samples being derivatives of the two earliest reconstructed cells in each mouse, and therefore that this first reconstructed cell division represented division of the fertilized egg (Fig. 2). It is also possible that daughter cells from earlier cell divisions did not contribute at all to the adult mice. If so, we would be unable to detect their existence.

Several features of normal mouse development can be extracted from these reconstructions. The two daughter cells of early embryonic cell divisions often contribute unequally to adult tissues. For example, one of the daughter cells in the first generation of reconstructed cell divisions in mouse 1 (cell 'b' in Fig. 1a) is the progenitor for 12 organoid clones, whereas the other is the progenitor of just one (cell 'c'). Examination of the tail from this animal, using the sequencing read counts of the mutations found in the first two daughter cells, again showed unequal contributions of the two progenitors: ~75% from one and ~30% from the other (cells 'b' and 'c' in Fig. 2a). A similar degree of inequality of contribution is shown for the first reconstructed cell division of mouse 2

[1]Cancer Genome Project, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK. [2]Department of Paediatrics, University of Cambridge, Hills Road, Cambridge CB2 2XY, UK. [3]Hubrecht Institute, Royal Netherlands Academy of Arts and Sciences, CancerGenomiCs.nl & University Medical Center Utrecht, 3584 CT, Utrecht, The Netherlands. [4]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK. [5]East Anglian Medical Genetics Service, Cambridge University Hospitals NHS Foundation Trust, Hills Road, Cambridge CB2 0QQ, UK. †Present address: Wellcome Trust/Cancer Research UK Gurdon Institute, Tennis Court Road, Cambridge CB2 1QN, UK.
*These authors contributed equally to this work.

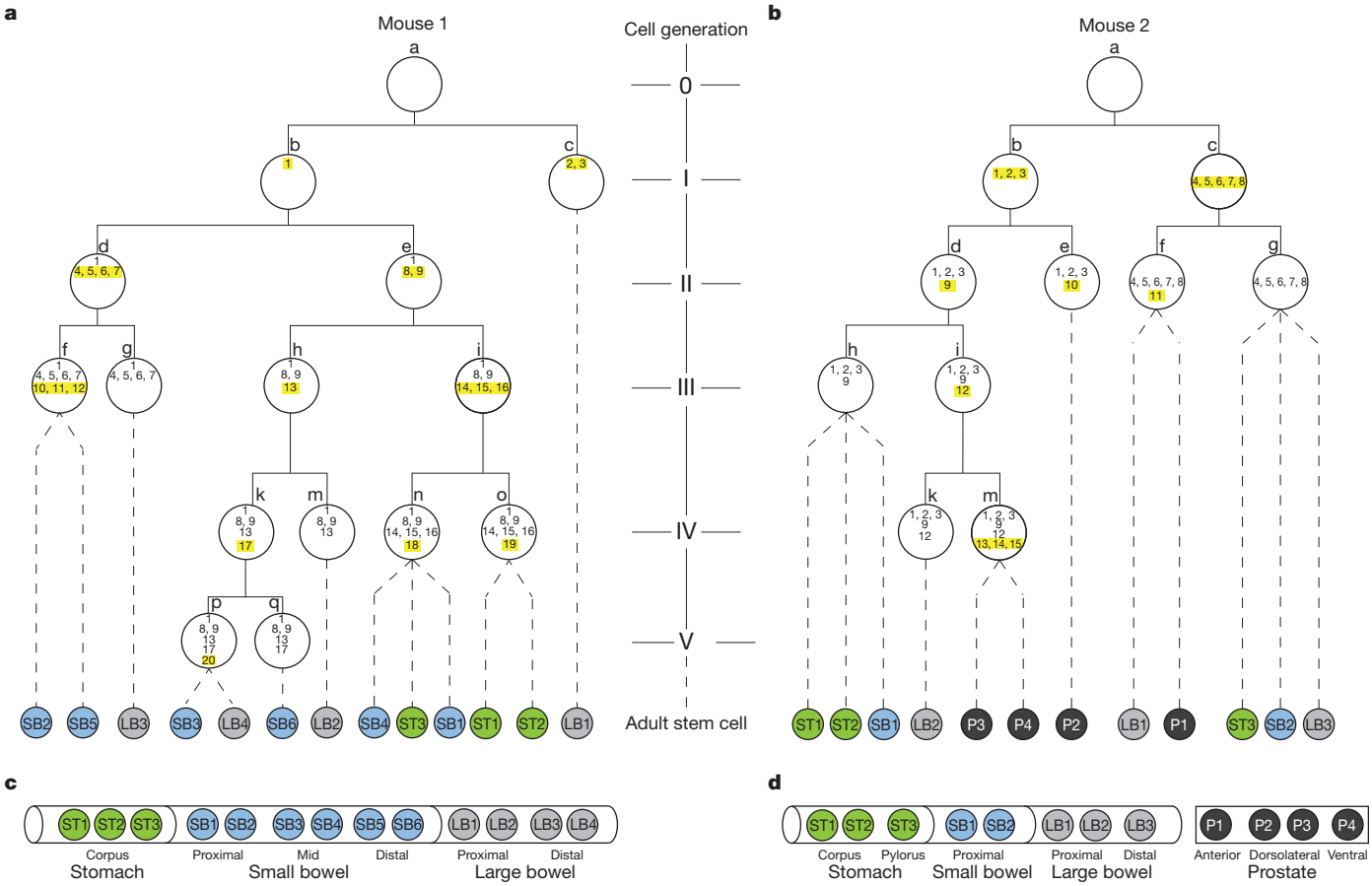


Figure 1 | Reconstructed phylogenetic trees of cells from early mouse embryos. **a**, Mouse 1. **b**, Mouse 2. Each white-filled large circle represents an embryonic cell that is defined by a unique combination of mutations. Each mutation is represented by a number inside the white circles. Yellow highlighted numbers: mutations acquired during most recent mitosis. Letters next to white circles: identifiers of each embryonic cell. Roman numerals indicate each reconstructed cell generation. **c**, **d**, Colour-filled smaller circles represent individual organoids derived from different anatomical regions of mouse 1 (**c**) and mouse 2 (**d**). Dashed lines connect each organoid with its last identifiable embryological precursor. An unknown number of cell generations lies between each organoid and its last identifiable embryological precursor.

(cells 'b' and 'c' in Fig. 2b). These observations are consistent with results obtained by other approaches[18,19]. This asymmetry propagates beyond the first cell division in both mice (Fig. 2).

At these early stages of embryo development, individual cells contribute to multiple tissues. For example, at least three of the four reconstructed cells from the two mice at generation I are ancestors of all tissues sampled, including the tail, and are thus probably contribute to endoderm, mesoderm and ectoderm (mouse 1, cell 'b'; mouse 2, cells 'b' and 'c'; Figs 1a, b and 2a, b). A similarly broad tissue contribution is made by at least one reconstructed cell that has undergone two cell divisions

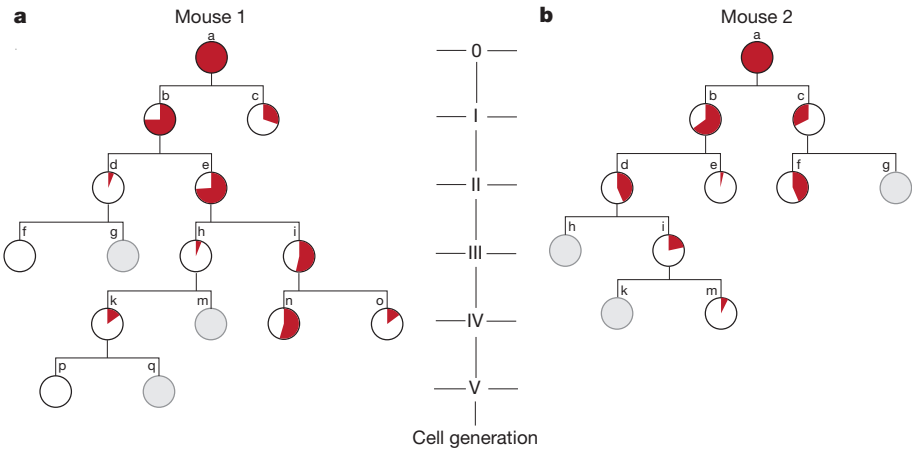


Figure 2 | Contributions of early embryonic cells to adult tail cell populations. Early embryonic phylogenetic trees of cells from mouse 1 (**a**) and mouse 2 (**b**), as in Fig. 1. The proportional contribution of each embryonic precursor cell to the population of cells in the tail is represented by the proportion of the circle area coloured red. This contribution was determined by assessment of the read count, in the tail, of the most recently acquired mutation(s) in each early embryo cell. Grey embryonic precursor cells did not acquire new mutations in the most recent mitosis, so their contributions to the tail cannot be measured directly.

(mouse 2, cell 'd'; Fig. 1b). Cell 'i', which has undergone three cell divisions in mouse 2, is the progenitor of prostate and large bowel cells and cells in tail tissues. The most distant early embryonic cells from the fertilized egg that we can reconstruct have undergone five cell divisions (mouse 1, cells 'p' and 'q'; Fig. 1a), and one of these contributes to both the distal colon and mid small intestine. Similarly, each organ derives from multiple early embryonic progenitor cells. For example, prostate in mouse 2 is contributed to by both reconstructed cells in generation I (cells 'b' and 'c' in Fig. 1b) and by at least three of the four cells in generation II (cells 'd', 'e' and 'f' in Fig. 1b).

Mutations in individual cells can also provide insights into the number of cell divisions and the mutational processes that have been operative in normal tissues. To explore these questions, we extracted 25 sets of somatic substitutions (6,714 in total) unique to individual clones (with a true positive rate of 92%, as assessed by validation of 743 variants, and a mean sensitivity of 50% for detecting heterozygous substitutions (Supplementary Table 1)). Of the four tissues sampled, small-bowel stem cells have acquired significantly more base substitutions than any other tissue, and prostate and stomach the fewest (Fig. 3b; see Methods for an explanation of statistical test and *P* values). This remains true if the analysis is restricted to C→T substitutions at CpG dinucleotides, which are predominantly due to a ubiquitous, intrinsic mutation process depending on the deamination of 5-methylcytosine to thymine[2]. These differences could be due to different mutation rates per cell division in different tissues, different numbers of cell divisions in the lineage from fertilized egg to adult stem cell in different tissues, or both. Using other approaches, it has been estimated that the rate of small-bowel stem cell division is greater than that of cells from colon or stomach[20], consistent with the mutation counts demonstrated here, suggesting that the number of somatic substitutions is acting as a 'cell division clock'. The murine small-bowel Lgr5-positive stem cell has been reported to divide every 21.5 h (ref. 21). Thus, the number of substitution mutations arising in each cell division in small-bowel stem cells is ~1.1 per cell division (Methods), similar to the rate of ~1.5 per cell division during early embryogenesis (Supplementary Discussion). The burden of mutations was higher in zones of repressed chromatin ($q < 0.00001$) and late-replicating regions ($P = 0.000034$), patterns previously reported for human cancer[22,23] (Extended Data Fig. 5). There was no statistically significant difference in mutation burden between the two mice.

Comparison of the six subclasses of base substitution between the four mouse tissues sampled revealed differences in mutation patterns. In particular, the proportion of C→A substitutions in small-bowel stem cells was greater than in other adult tissues and in the 35 aggregated early embryonic mutations (Fig. 3a, c), which were predominantly C→T transitions enriched at CpG dinucleotides. Analysis by non-negative matrix factorization[24], incorporating the immediate 5′ and 3′ sequence context

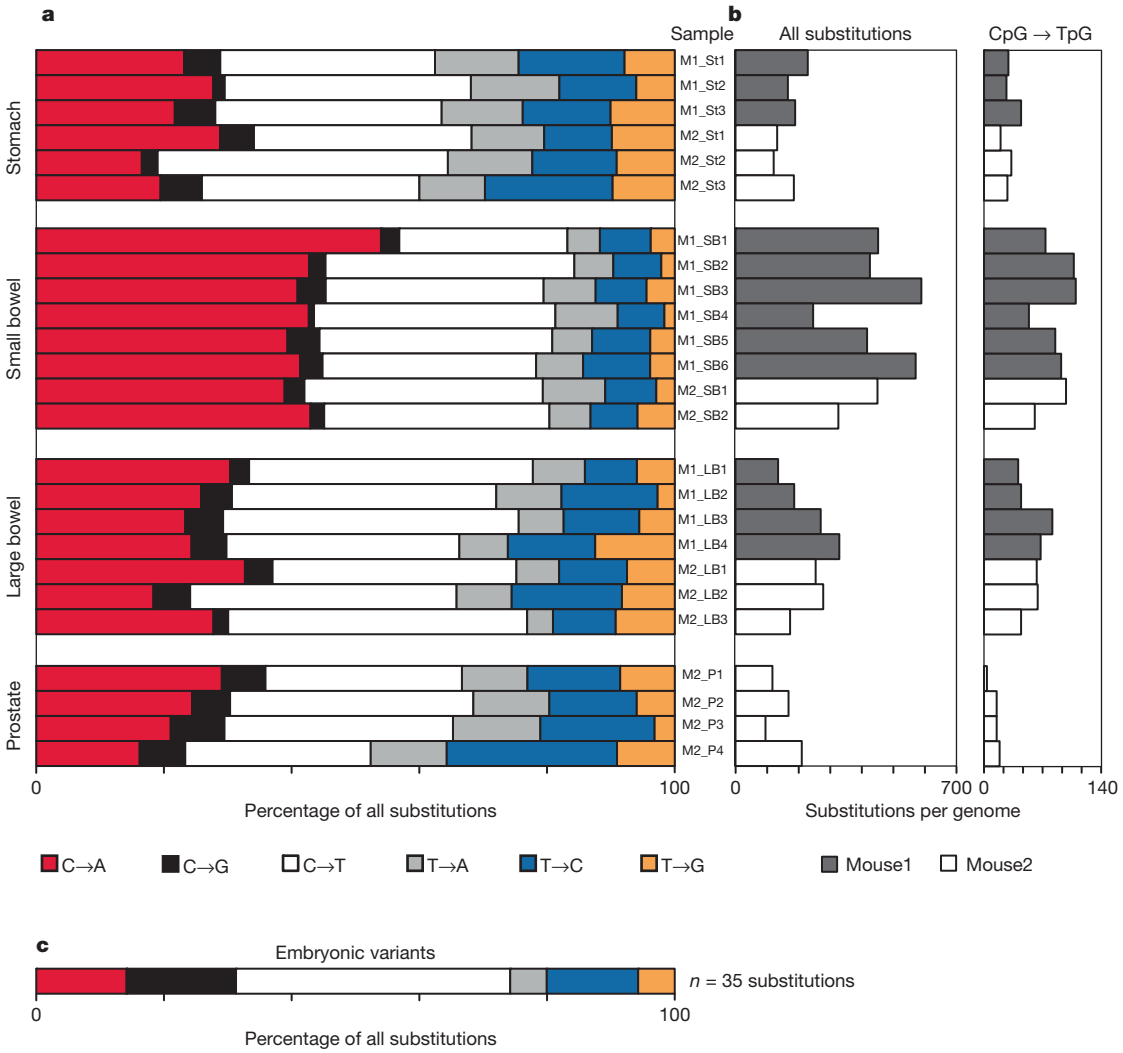


Figure 3 | The number and spectrum of substitution mutations in individual organoids. **a**, The spectrum of base substitution mutations in each organoid. **b**, Left: the absolute number of genome-wide substitutions. When adjusted for sensitivity, the range of the mutation burden per organoid genome is 179–1,190 (mean 609) substitutions (see Supplementary Table 1). Right: the number of CpG→TpG substitutions. Grey bars, mouse 1; white bars, mouse 2. **c**, Mutational spectrum of embryonic variants.

of each mutated base, extracted two underlying mutational signatures. One was characterized predominantly by C→T mutations (signature 2 in Extended Data Fig. 6). The other had a major feature of C→A substitutions at XpCpT trinucleotides in addition to C→T mutations (signature 1 in Extended Data Fig. 6). Mutations acquired *in vitro* by small-bowel organoid stem cells had a different mutational signature. Sequencing of subclones of cell populations expanded from two small bowel single stem cells that had been exposed to a period of culture revealed a distinct mutational signature characterized by T→G mutations enriched at XpTpT trinucleotides (Extended Data Fig. 7). The data indicate that multiple mutational processes are operative in normal cells *in vivo* and *in vitro* and that the degree of exposure differs across tissues. The mechanism underlying the signature characterized by C→A mutations is unclear. However, one possible cause is mutagenesis by reactive oxygen species, which have been reported to generate C→A mutations[25].

Catalogues of somatic mutations act as 'archaeological' records reflecting the past experiences of cancer cells[26]. Our results illustrate the insights that they can provide into the life histories of normal cells. Sequencing of larger numbers of individual cells from a wider range of tissues will allow the precise reconstruction of cell lineages extending into later stages of embryogenesis with further insights into tissue-specific development, mutational processes and mutation rates. Studies of multiple animals will reveal variability between individuals and relationships with age, disease and environmental exposures. If somatic mutation rates in humans are similar to those in mice, the results indicate that application of this approach to human cells is a tractable endeavour.

METHODS SUMMARY

Two homozygous C57Bl6 male mice (mouse 1 and mouse 2) with *Lgr5-ki*, derived from lines established in 2006/2007, were euthanized at 116 and 98 weeks of age, respectively. Clonal organoid lines were grown from single clonal glands (stomach)[13,17,27], single clonal crypts (small intestine[16,17] and colon[15,17]) or single cells (prostate; W.K. and H.C., manuscript in preparation). DNA from organoids and tail tissue from each mouse were whole-genome sequenced[28]. Somatic single-nucleotide substitutions unique to each organoid were called using the CaVEMan algorithm[28], comparing each organoid with the tail of the same mouse. The precision of these mutations was 92%, as determined by massively parallel sequencing of PCR amplicons targeting 743 (11%) randomly selected substitutions. Early embryonic mutations may be shared by multiple organoids and thus can be distinguished from mutations occurring later in development or after birth. However, they may also be present in the tail of the same mouse at significant read counts and thus be removed in comparisons of organoids against the tail. To derive comprehensive catalogues of embryonic mutations, we therefore compared each organoid against the tail of the other mouse. Every potential embryonic variant thus detected was subjected to validation by capillary sequencing in every organoid and control tissue. We constructed lineage trees from mutations that were confirmed to be somatic and shared by at least two organoids.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 11 December 2013; accepted 7 May 2014.
Published online 29 June 2014.

1. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Rev. Genet.* **14,** 618–630 (2013).
2. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).
3. Sato, T. & Clevers, H. Growing self-organizing mini-guts from a single intestinal stem cell: mechanism and applications. *Science* **340,** 1190–1194 (2013).
4. De, S. Somatic mosaicism in healthy human tissues. *Trends Genet.* **27,** 217–223 (2011).
5. Carlson, C. A. *et al.* Decoding cell lineage from acquired mutations using arbitrary deep sequencing. *Nature Methods* **9,** 78–80 (2012).
6. Kennedy, S. R., Loeb, L. A. & Herr, A. J. Somatic mutations in aging, cancer and neurodegeneration. *Mech. Ageing Dev.* **133,** 118–126 (2012).
7. Salipante, S. J. & Horwitz, M. S. Phylogenetic fate mapping. *Proc. Natl Acad. Sci. USA* **103,** 5448–5453 (2006).
8. Salipante, S. J. & Horwitz, M. S. A phylogenetic approach to mapping cell fate. *Curr. Top. Dev. Biol.* **79,** 157–184 (2007).
9. Shibata, D. & Tavare, S. Counting divisions in a human somatic cell tree: how, what and why? *Cell Cycle* **5,** 610–614 (2006).
10. Wasserstrom, A. *et al.* Reconstruction of cell lineage trees in mice. *PLoS ONE* **3,** e1939 (2008).
11. Zhou, W. *et al.* Use of somatic mutations to quantify random contributions to mouse development. *BMC Genomics* **14,** 39 (2013).
12. Lasken, R. S. Single-cell sequencing in its prime. *Nature Biotechnol.* **31,** 211–212 (2013).
13. Barker, N. *et al.* Lgr5^{+ve} stem cells drive self-renewal in the stomach and build long-lived gastric units *in vitro*. *Cell Stem Cell* **6,** 25–36 (2010).
14. Barker, N. *et al.* Identification of stem cells in small intestine and colon by marker gene *Lgr5*. *Nature* **449,** 1003–1007 (2007).
15. Sato, T. *et al.* Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. *Gastroenterology* **141,** 1762–1772 (2011).
16. Sato, T. *et al.* Single Lgr5 stem cells build crypt-villus structures *in vitro* without a mesenchymal niche. *Nature* **459,** 262–265 (2009).
17. Snippert, H. J. *et al.* Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. *Cell* **143,** 134–144 (2010).
18. Plusa, B. *et al.* The first cleavage of the mouse zygote predicts the blastocyst axis. *Nature* **434,** 391–395 (2005).
19. Bruce, A. W. & Zernicka-Goetz, M. Developmental control of the early mammalian embryo: competition among heterogeneous cells that biases cell fate. *Curr. Opin. Genet. Dev.* **20,** 485–491 (2010).
20. Barker, N. *et al.* Very long-term self-renewal of small intestine, colon, and hair follicles from cycling *Lgr5*$^{+ve}$ stem cells. *Cold Spring Harb. Symp. Quant. Biol.* **73,** 351–356 (2008).
21. Schepers, A. G., Vries, R., van den Born, M., van de Wetering, M. & Clevers, H. Lgr5 intestinal stem cells have high telomerase activity and randomly segregate their chromosomes. *EMBO J.* **30,** 1104–1109 (2011).
22. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).
23. Schuster-Bockler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488,** 504–507 (2012).
24. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3,** 246–259 (2013).
25. van Loon, B., Markkanen, E. & Hubscher, U. Oxygen as a friend and enemy: how to combat the mutational potential of 8-oxo-guanine. *DNA Repair (Amst.)* **9,** 604–616 (2010).
26. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149,** 994–1007 (2012).
27. Leushacke, M., Ng, A., Galle, J., Loeffler, M. & Barker, N. Lgr5^{+} gastric stem cells divide symmetrically to effect epithelial homeostasis in the pylorus. *Cell Rep.* **5,** 349–356 (2013).
28. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149,** 979–993 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank M. Zernicka-Goetz and the Goldman group for discussion of our findings. This work was supported by funding from the Wellcome Trust (grant reference 077012/Z/05/Z), the Kadoorie Charitable Foundation and the Louis-Jeantet Foundation. Individual authors were supported as follows: M.H., Marie Curie IEF fellowship (EU/236954) and ERC grant (232814); R.B. and E.C., Zenith grant of the Netherlands Genomics Initiative (935.12.003); W.K., Centre for Biomedical Genetics, Utrecht; I.M., EMBO Long Term Fellowship (ALTF-1287-2012); S.N.Z., Wellcome Trust Intermediate Clinical Fellowship (WT100183MA) and Wellcome-Beit Prize Fellowship 2013; S.B., Wellcome Trust Research Training Fellowship for Clinicians; and P.C., Wellcome Trust Senior Research Fellowship in Clinical Science.

Author Contributions S.B. and M.R.S. analysed sequencing data. R.B. and E.C. contributed data and data analyses. S.R., M.P. and P.S.T. contributed to data analysis. I.M. assessed the association of mutation density with genomic features. L.A. performed analysis of mutational signatures. D.W. and P.C. performed statistical analyses. S.B, S.N.Z., P.C. and M.R.S. contributed to data interpretation. M.H., W.K. and M.W. generated organoids. M.M., L.M. and B.R. performed technical investigations. A.T. and N.G. performed phylogenetic analyses. H.C. and M.R.S. directed the research. M.R.S. wrote the manuscript.

Author Information Sequencing data have been deposited at the European Nucleotide Archive (http://www.ebi.ac.uk/ena/) under accession numbers ERP002057 (Illumina data) and ERP005717 (SOLiD data). Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the paper on www.nature.com/nature. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.R.S. (mrs@sanger.ac.uk).

LETTER

doi:10.1038/nature13725

Bidirectional switch of the valence associated with a hippocampal contextual memory engram

Roger L. Redondo[1,2]*, Joshua Kim[1]*, Autumn L. Arons[1,2], Steve Ramirez[1], Xu Liu[1,2] & Susumu Tonegawa[1,2]

The valence of memories is malleable because of their intrinsic reconstructive property[1]. This property of memory has been used clinically to treat maladaptive behaviours[2]. However, the neuronal mechanisms and brain circuits that enable the switching of the valence of memories remain largely unknown. Here we investigated these mechanisms by applying the recently developed memory engram cell-manipulation technique[3,4]. We labelled with channelrhodopsin-2 (ChR2) a population of cells in either the dorsal dentate gyrus (DG) of the hippocampus or the basolateral complex of the amygdala (BLA) that were specifically activated during contextual fear or reward conditioning. Both groups of fear-conditioned mice displayed aversive light-dependent responses in an optogenetic place avoidance test, whereas both DG- and BLA-labelled mice that underwent reward conditioning exhibited an appetitive response in an optogenetic place preference test. Next, in an attempt to reverse the valence of memory within a subject, mice whose DG or BLA engram had initially been labelled by contextual fear or reward conditioning were subjected to a second conditioning of the opposite valence while their original DG or BLA engram was reactivated by blue light. Subsequent optogenetic place avoidance and preference tests revealed that although the DG-engram group displayed a response indicating a switch of the memory valence, the BLA-engram group did not. This switch was also evident at the cellular level by a change in functional connectivity between DG engram-bearing cells and BLA engram-bearing cells. Thus, we found that in the DG, the neurons carrying the memory engram of a given neutral context have plasticity such that the valence of a conditioned response evoked by their reactivation can be reversed by re-associating this contextual memory engram with a new unconditioned stimulus of an opposite valence. Our present work provides new insight into the functional neural circuits underlying the malleability of emotional memory.

The amygdala can encode both negative and positive valence[5–10] and the DG encodes contextual information[3,11,12]. It is unknown whether the DG drives expression of memories irrespective of the valence of the unconditioned stimulus (US). Therefore we investigated the roles of DG and BLA engrams in determining the valence of contextual memories and its possible switch. We targeted engram-bearing cells by infecting DG and BLA neurons of *c-fos*-tTA male mice with AAV_9 virus expressing, under the control of the tetracycline response element (TRE), ChR2 and mCherry fusion protein (DG and BLA ChR2 mice, respectively) or mCherry-only (DG mCherry-only and BLA mCherry-only mice, respectively) (Fig. 1a, b and Methods). This method targets the expression of ChR2 to neurons in which the immediate early gene *c-fos* is activated during the encoding of a memory in the absence of the antibiotic doxycycline (dox) in the diet[3,13]. A similar proportion of neurons expressed ChR2 after encoding a fear memory (foot shock) or a reward memory (interaction with a female mouse in the home cage) (Fig. 1c). To test the capability of the ChR2-labelled neurons to drive an aversive or appetitive response, we developed two real-time optogenetic place memory tests: the optogenetic place avoidance (OptoPA) test for assessing aversive behaviour and the optogenetic place preference (OptoPP) test for assessing appetitive behaviour (see Methods and the characterization of these tests in Extended Data Fig. 1).

On day 1 of the protocol, the mice underwent a habituation session while on doxycycline in the OptoPA or OptoPP test (Fig. 1d, f). During habituation, light activation had similar effects on both the ChR2-expressing mice and the mCherry-only controls (Fig. 1e, g) and was not sufficient to produce a change in preference (Extended Data Fig. 2). On day 3, off doxycycline, mice habituated to the OptoPA test were fear conditioned in context A (fear memory group), whereas the mice habituated to the OptoPP test were reward conditioned, by spending 2 h in context B with one female mouse (reward memory group). As a negative control, groups of DG ChR2 or BLA ChR2 mice did not receive the US (foot shock in context A or female exposure in context B) on day 3 (DG ChR2, no US on day 3 or BLA ChR2, no US on day 3 mice). At the end of day 3, all animals were returned to a doxycycline diet, closing the time window for labelling for the remainder of the experiment. On day 5, both DG ChR2 and BLA ChR2 mice of the fear memory group exhibited greater aversive responses in the OptoPA test than the corresponding mCherry-only and ChR2, no US on day 3 mice (Fig. 1e). Both DG ChR2 and BLA ChR2 mice of the reward memory group showed greater appetitive response than corresponding mCherry-only mice (Fig. 1g), DG ChR2, no US on day 3 mice or BLA ChR2, no US on day 3 mice in the OptoPP test (Fig. 1g). Therefore, the US is necessary on day 3 for the engram neurons to drive the appropriate response on day 5 tests.

To investigate whether the valence of the memory associated with the DG or BLA engram can be reversed, we conducted within-subject longitudinal experiments. For this purpose, both the fear and reward memory groups were returned to a doxycycline diet immediately after day 3 conditioning, preventing the expression of ChR2 by other neurons that may upregulate *c-fos* promoter-driven ChR2 after day 3. In a 'fear-to-reward' experiment (Fig. 2a, b), the fear memory group was subjected to an OptoPA test on day 5, and as expected, DG-ChR2 and BLA-ChR2 mice exhibited aversive responses that were greater than DG mCherry-only mice and BLA mCherry-only mice, respectively (Fig. 2b). On day 7, these mice received light stimulation while interacting with 2 female mice in their home cage. This procedure on day 7 is hereafter referred to as 'induction'. Another group of DG ChR2 mice received light stimulation but no female mice (DG ChR2, no US on day 7 mice). On day 9, the OptoPP test was used in the fear-to-reward experiment to test whether the neurons activated by the light stimulation during the induction procedure could now drive an appetitive response. The DG ChR2 mice showed a greater appetitive response than the DG mCherry-only or DG ChR2, no US on day 7 mice (Fig. 2b). Light reactivation of the original fear memory engram-bearing cells labelled in the BLA (BLA ChR2 mice) failed to exhibit an appetitive response in the day 9 OptoPP test. For the fear-to-reward experiment, we introduced another protocol for DG ChR2 and BLA ChR2 mice consisting of OptoPA tests on day 5 and day 9. The DG ChR2 mice displayed lower aversive responses on day 9 than the day 5 test, whereas the DG ChR2, no US on day 7 mice and the BLA ChR2

[1]RIKEN–MIT Center for Neural Circuit Genetics at the Picower Institute for Learning and Memory, Department of Biology and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [2]Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.
*These authors contributed equally to this work.

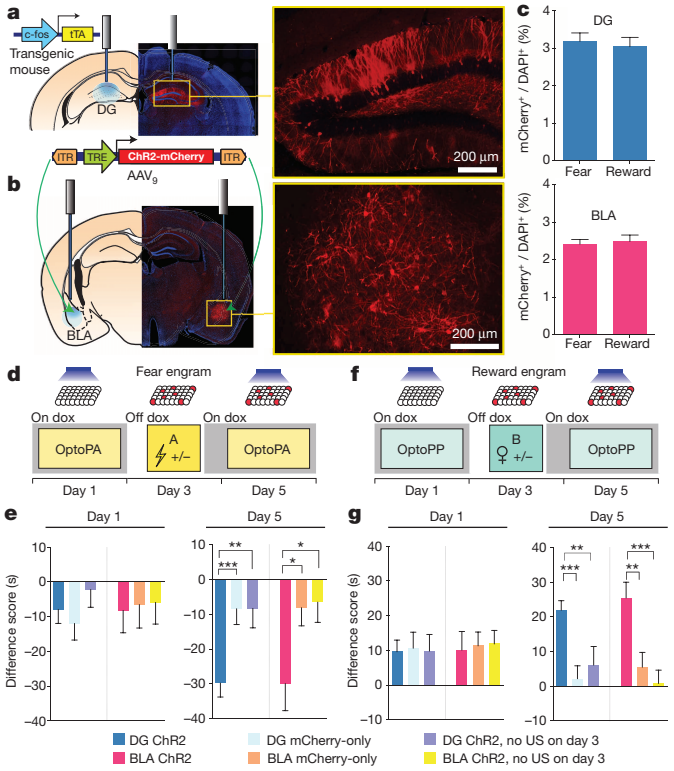


Figure 1 | Fear and reward engram reactivation, both in the DG and the BLA, drives place avoidance and place preference, respectively. **a**, **b**, *c-fos-tTA* mice were injected with AAV_9–TRE–ChR2–mCherry or TRE–mCherry and implanted with optical fibres bilaterally targeting the DG (**a**) or the BLA (**b**). **c**, Similar engram labelling in the DG ($t_{33} = 0.42$, NS (not significant)) and BLA ($t_{26} = 0.35$, NS) after fear (DG $n = 16$; BLA $n = 14$) and reward (DG $n = 19$; BLA $n = 14$) conditioning. **d**, Fear memory group experimental protocol. **e**, On day 1, difference scores (time in target zone during the on phase minus time in target zone during baseline phase) (Extended Data Fig. 1) were similar across all DG subgroups ($F_{2,101} = 0.76$, NS) and across all BLA subgroups ($F_{2,72} = 0.03$, NS). On day 5, difference scores were lower in DG ChR2 ($n = 48$) and BLA ChR2 mice ($n = 21$) compared to the corresponding mCherry-only (DG $n = 39$; BLA $n = 27$) and DG or BLA ChR2, no US on day 3 mice (DG $n = 17$; BLA $n = 27$) (DG $F_{2,101} = 7.99$, $P < 0.001$; BLA $F_{2,72} = 4.12$, $P < 0.05$). **f**, Reward memory group experimental protocol. **g**, On day 1, difference scores were similar across all DG subgroups ($F_{2,111} = 0.02$, NS) and across all BLA subgroups ($F_{2,83} = 0.04$, NS). In the day 5 OptoPP test, difference scores were greater in DG ChR2 ($n = 54$) and BLA ChR2 mice ($n = 35$) compared to corresponding mCherry-only (DG $n = 36$; BLA $n = 31$) and DG or BLA ChR2, no US on day 3 mice (DG $n = 24$; BLA $n = 21$) (DG $F_{2,111} = 9.76$, $P < 0.001$; BLA $F_{2,83} = 9.12$, $P < 0.001$). Significance for multiple comparisons: $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$. Results show mean ± s.e.m.

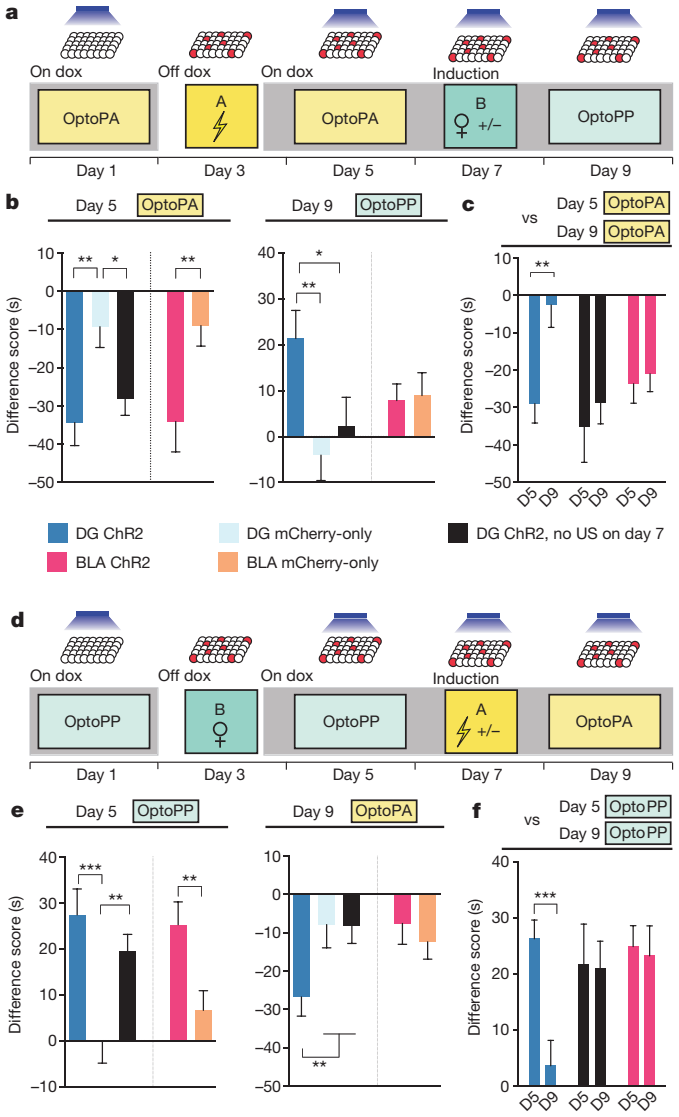


Figure 2 | The valence associated with the DG engram is reversed after induction with the unconditioned stimulus of opposite value. **a**, The fear-to-reward experimental protocol. **b**, On day 5, difference scores of DG ChR2 ($n = 16$); DG ChR2, no US on day 7 ($n = 20$); and BLA ChR2 mice ($n = 19$) were lower compared to corresponding mCherry-only mice (DG $n = 27$; BLA $n = 27$) (DG $F_{2,59} = 6.16$, $P < 0.01$; BLA $t_{44} = 2.73$, $P < 0.01$). In the day 9 OptoPP test, difference scores of DG ChR2 mice were greater than the control mice ($F_{2,60} = 4.4$, $P < 0.05$). Difference scores of BLA ChR2 mice were similar to those of BLA mCherry-only mice ($t_{44} = 0.16$, NS). **c**, On the day 9 OptoPA test, DG ChR2 mice ($n = 12$) showed a less aversive response compared to day 5, whereas both DG ChR2, no US on day 7 ($n = 16$) and BLA ChR2 ($n = 17$) mice showed similar difference scores on these days ($F_{1,42} = 5.42$, $P < 0.05$). **d**, Reward-to-fear experimental protocol. **e**, On day 5 OptoPP test, difference scores of DG ChR2 ($n = 17$); DG ChR2, no US on day 7 ($n = 29$); and BLA ChR2 mice ($n = 30$) were greater compared to the corresponding mCherry-only mice (DG $n = 27$; BLA $n = 29$) (DG $F_{2,70} = 8.97$, $P < 0.001$; BLA $t_{57} = 2.85$, $P < 0.01$). On day 9, difference scores of DG ChR2 mice were lower compared to the control mice ($F_{2,71} = 3.20$, $P < 0.05$) and difference scores of BLA ChR2 mice were similar to those of BLA mCherry-only mice ($t_{57} = 0.49$, NS). **f**, On the day 9 OptoPP test, DG ChR2 mice ($n = 18$) showed a reduced appetitive response compared to day 5 OptoPP test, whereas both DG ChR2, no US on day 7 ($n = 21$) and BLA ChR2 mice ($n = 18$) showed similar difference scores on day 9 and day 5 ($F_{1,54} = 6.58$, $P < 0.05$). Significance for multiple comparisons: $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$. Results show mean ± s.e.m.

mice showed similar aversive responses on day 5 and day 9 (Fig. 2c), indicating that the DG ChR2 mice not only acquired reward memory by the induction procedure but also lost much of the fear memory acquired previously. In contrast, BLA ChR2 mice neither acquired a reward memory nor lost the fear memory.

In the reward-to-fear experiment (Fig. 2d, e), DG ChR2 and BLA ChR2 mice as well as the DG ChR2, no US on day 7 mice displayed greater appetitive responses in the OptoPP test on day 5 than the corresponding mCherry-only mice (Fig. 2e). On day 7, these mice received light stimulation either while being subjected to fear conditioning in context A (DG ChR2 and BLA ChR2 mice) or while exploring context A without foot shocks (DG ChR2, no US on day 7 mice). DG ChR2 mice, but not DG ChR2, no US on day 7 mice displayed a greater aversive response in the day 9 OptoPA test than DG mCherry-only mice (Fig. 2e). Among several groups of mice, only in the DG ChR2 mice did the same neuronal ensemble whose reactivation led to an appetitive response

during the OptoPP test on day 5 produced an aversive response on day 9. For the reward-to-fear experiment, we introduced another protocol for DG ChR2 and BLA ChR2 mice consisting of OptoPP tests on day 5 and day 9. The DG ChR2 mice displayed lower appetitive responses on day 9 than the day 5 OptoPP test, whereas the DG ChR2, no US on day 7 mice and the BLA ChR2 mice showed similar appetitive responses on day 5 and day 9 (Fig. 2f), mirroring the results obtained by the fear-to-reward experiment described above (Fig. 2c).

We next investigated the effect of our manipulations at the cellular level (Fig. 3a, b; see Methods for the experimental design details). Three groups were defined by whether, on day 3 of the protocol, mice experienced an event of the opposite valence without engram reactivation (light$^-$ US$^+$), induction (light$^+$ US$^+$), or received optogenetic stimulation but no US delivery (light$^+$ US$^-$). On day 5, the effect of DG engram reactivation on the BLA was assessed. The proportions of cells (mCherry$^+$ / DAPI$^+$) (Fig. 3c, d) labelled on day 1 as well as the proportion of cells activated on day 5 (GFP$^+$ / DAPI$^+$)(Fig. 3e, f) were similar across experimental groups in both the DG and BLA. The proportion of DG engram cells labelled on day 1 that were light-reactivated on day 5 (GFP$^+$mCherry$^+$ / mCherry$^+$) was similar in all groups (Fig. 3g) and above chance (Fig. 3h). In the light$^-$ US$^+$ and the light$^+$ US$^-$ groups, the proportion of BLA engram cells labelled on day 1 that were reactivated on day 5 (GFP$^+$mCherry$^+$ / mCherry$^+$) by artificial reactivation of DG engram cells were similar (Fig. 3i, k) and significantly greater than chance (Fig. 3j). In the light$^+$ US$^+$ group, this proportion, though greater than chance overlap (Fig. 3j), was significantly lower compared to the light$^-$ US$^+$ and the light$^+$ US$^-$ groups (Fig. 3i, k). This suggests that the induction procedure decreased the ability of the DG engram to activate the BLA engram, indicating a change in their functional connectivity[14,15].

Using the reward-to-fear scheme, we next investigated the effect of the light reactivation of a previously labelled (that is, day 3) reward memory engram on subsequent encoding of a fear memory (that is, day 7) and whether this procedure affects the recall of the fear memory by natural cues (day 11) (Fig. 4a). On day 7, during induction, DG ChR2 and BLA ChR2 mice displayed significantly lower freezing than DG mCherry-only and BLA mCherry-only mice, respectively (Fig. 4b). In addition, DG ChR2 and BLA ChR2 mice displayed significantly lower freezing compared to DG ChR2, no US on day 3 and BLA Ch2, no US on day 3 mice, respectively on day 7 and day 11 (Fig. 4b), indicating that the reduced encoding and recall of the fear memory in DG ChR2 and BLA ChR2 depends on the rewarding experience on day 3. Finally, we studied the integrity of the originally acquired memory after the induction of a memory of the opposite valence. For this purpose, in the fear-to-reward scheme, we investigated the effect of the reward memory induction on day 7 on the original fear memory by testing the freezing response to the original context on day 11 (Fig. 4c). DG ChR2 mice, but not DG mCherry-only and DG ChR2, no US on day 7 mice, showed a significant reduction in their freezing response even though these mice were never re-exposed to the original context between encoding of the original fear memory on day 3 and testing on day 11 (Fig. 4d). In contrast, BLA ChR2, but not BLA mCherry-only mice, showed elevated freezing during the day 11 test session compared to the encoding session on day 3, consistent with earlier observations that BLA engram reactivation leads to the sensitization of fear responses[16,17].

During the context A test on day 11, DG ChR2 mice spent more time sniffing, a behavioural response that increases in the presence of female cues and decreases after fear training[18,19] (corrected for freezing periods) (see Methods), than any other groups of mice (Fig. 4e), suggesting that our reward memory induction procedure increased the positive memory valence associated with the test context (that is, context A) on day 11.

Here we have shown that both the DG and BLA neurons activated during context-specific fear or reward conditioning can drive aversive and appetitive responses, respectively, upon optogenetic reactivation of these cells two days after training. This confirms our previous finding

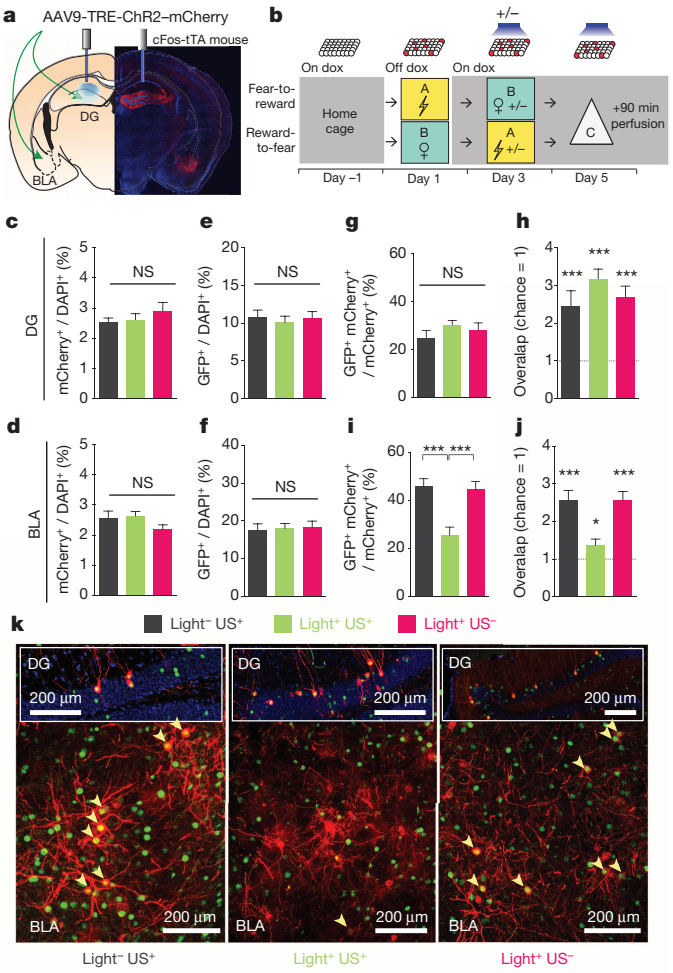


Figure 3 | DG to BLA functional connectivity changes after induction. **a**, Injection sites and the optic fibre placement. **b**, Experimental protocol. On day 1, cells active during fear or reward experience were labelled. On day 3, on doxycycline, mice were randomly assigned to 3 groups: reward or fear conditioning without light reactivation (light$^-$ US$^+$) ($n = 11$); full induction protocol (light$^+$ US$^+$) ($n = 16$); light stimulation but neither reward nor fear conditioning (light$^+$ US$^-$) ($n = 16$). On day 5, all animals received light stimulation in the DG for 12 min in a novel context (context C) before the brains were used for immunohistochemistry. **c**, **d**, Similar proportions of neurons were labelled by the fear or reward conditioning on day 3 in all groups, both in DG ($F_{2,40} = 0.77$, NS) (**c**) and BLA ($F_{2,40} = 2.40$, NS) (**d**). **e**, **f**, Light delivery to the DG on day 5 led to the activation (GFP$^+$ / DAPI$^+$) of similar proportions of cells in DG ($F_{2,40} = 0.21$, NS) (**e**) and BLA ($F_{2,40} = 0.06$, NS) (**f**). **g**, **h**, Levels of reactivation (GFP$^+$mCherry$^+$ / mCherry$^+$) in the DG were similar across all groups ($F_{2,40} = 0.61$, NS) (**g**) and above levels of chance (one sample *t*-tests against chance overlap: light$^-$,US$^+$ $t_{10} = 4.24$, $P < 0.01$; light$^+$ US$^+$ $t_{15} = 8.56$, $P < 0.001$; Light$^+$ US$^-$ $t_{15} = 5.5$, $P < 0.001$) (**h**). **i**, **j**, Levels of reactivation (GFP$^+$ mCherry$^+$ / mCherry$^+$) in the BLA were lower in the light$^+$ US$^+$ compared to light$^-$ US$^+$ and light$^+$ US$^-$ ($F_{2,40} = 11.82$, $P < 0.001$), even though overlap levels (**j**) remained above chance (one sample *t*-tests: light$^-$ US$^+$ $t_{10} = 7.41$, $P < 0.001$; light$^+$ US$^+$ $t_{15} = 2.33$, $P < 0.05$; light$^+$ US$^-$ $t_{15} = 6.94$, $P < 0.001$). **k**, Representative images of double immunofluorescence for GFP (green) and mCherry (red) in the DG and BLA. Significance for multiple comparisons: $^*P < 0.05$; $^{***}P < 0.001$. Results show mean ± s.e.m.

that these neurons have undergone enduring changes as a consequence of memory training, validating their engram-bearing nature. We have also shown that artificially reactivating contextual fear-labelled neurons in the DG, but not in the BLA, during a subsequent reward conditioning was sufficient to reverse the dominant valence associated with the memory. Reciprocally, artificially reactivating contextual reward-labelled neurons

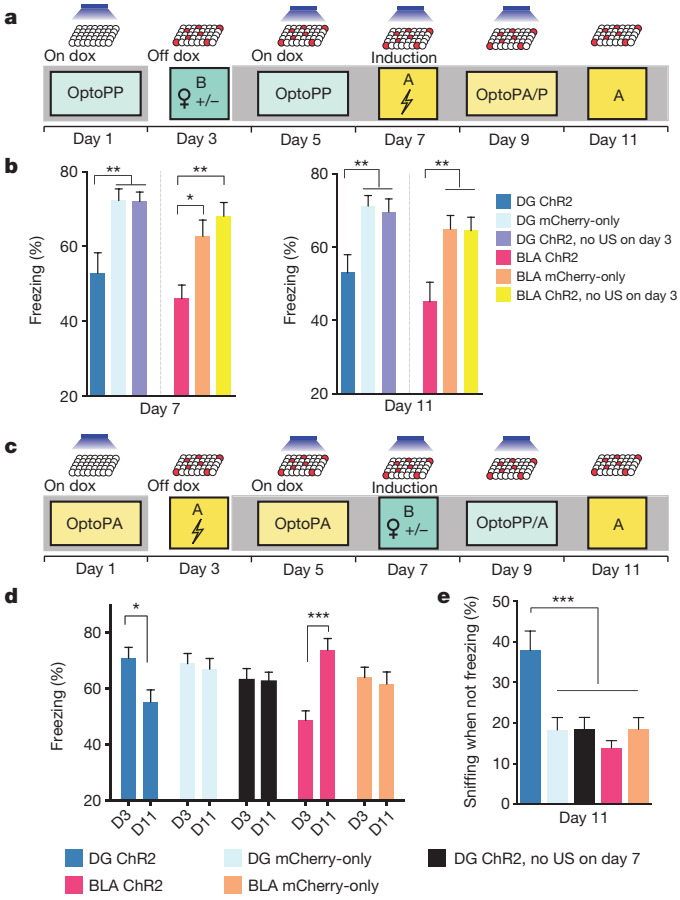


Figure 4 | Memory induction alters naturally cued fear memory. **a**, Reward-to-fear scheme with context A fear memory tests on day 7 and day 11. **b**, On day 7, freezing levels during the last 3 min of the induction procedure were reduced in DG ChR2 ($n = 18$) and BLA ChR2 mice ($n = 19$) compared to corresponding mCherry-only (DG $n = 21$; BLA $n = 20$) and ChR2, no US on day 3 mice (DG $n = 27$; BLA $n = 24$) (DG $F_{2,63} = 8.768$, $P < 0.001$; BLA $F_{2,60} = 8.49$, $P < 0.001$). These reduced freezing levels remained on day 11 (DG $F_{2,63} = 6.25$, $P < 0.01$; BLA $F_{2,60} = 6.86$, $P < 0.01$). **c**, Fear-to-reward scheme with context A fear memory tests on day 3 and day 11. **d**, Compared to the last three minutes of the fear conditioning on day 3, only DG ChR2 mice ($n = 27$) showed a reduction of freezing responses on day 11 (interaction $F_{4,151} = 8.48$, $P < 0.001$). BLA ChR2 mice ($n = 29$) showed increased levels of freezing on day 11 compared to day 3. DG ChR2, no US on day 7 mice ($n = 42$) and mCherry-only mice (DG $n = 32$; BLA $n = 29$) did not show differences between day 3 and day 11. **e**, After correcting for the time spent freezing, DG ChR2 mice spent a larger proportion of time sniffing than any other group in context A on day 11 ($F_{4,151} = 7.78$, $P < 0.001$). Significance for multiple comparisons: $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$. Results show mean ± s.e.m.

in the DG, but not in the BLA, during a subsequent contextual fear conditioning was sufficient to reverse the dominant valence associated with the original memory (Fig. 2). We conclude that the valence associated with the hippocampal memory engram is bidirectionally reversible. In contrast, the inability of the BLA engram to reverse the valence of the memory suggests that individual BLA cells are hardwired to drive either fear or reward memories rather than both.

The reversal of the dominant valence associated with the DG memory engram was also demonstrated at the cellular level. The hippocampal output has been shown to be sufficient to induce synaptic plasticity in the amygdala[20], and post-training inactivation of the dorsal hippocampus prevents context-dependent neuronal activity in the amygdala[21]. In addition, studies have shown that the BLA can drive both aversive[22] and appetitive[23] responses. We thus predicted that driving the DG engram associated with a particular valence would result in firing of the corresponding BLA engram-bearing cells active during encoding—a hypothesis that was supported by the data in Fig. 3 (light^{-} US^{+}). We hypothesized that optogenetic reactivation of the DG engram-bearing cells during the presentation of a US with a valence opposite to the original one would strengthen the connectivity, albeit indirectly, of these DG cells with a new subset of BLA neurons while weakening the connections established during the original learning. This hypothesis is supported by the finding that the overlap of BLA neurons activated by stimulation of the DG engram-bearing cells was reduced after induction compared to the overlap observed in no induction controls (Fig. 3i, k). The observation that the levels of *c-fos* activation were similar across groups suggests that a new population of BLA neurons was functionally recruited in the light^{+} US^{+} group.

By applying optogenetic manipulations to two interacting brain areas, our study provides a new type of neural circuit analysis that elucidates functional relationships between brain areas with respect to the expression of memories. Previously, others have shown that neurons that artificially upregulate CREB[24] and express TRPV1[16] in the BLA can be associated with fear. Also, randomly labelled populations of neurons in the piriform cortex can drive opposing behaviours after their stimulation is paired with a US of positive or negative valence[25]. Here, because our engram labelling technology allows for the targeting of neurons that were activated during natural memory encoding to express ChR2, we were able to label a specific memory trace and monitor the behavioural response elicited by natural cues (Fig. 4a, b). Our present study provides a new insight into the functional neural circuit underlying the malleability of emotional memory by highlighting the importance of the plasticity in the hippocampal–amygdala connections.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 19 December 2013; accepted 30 July 2014.
Published online 27 August 2014.

1. Pavlov, I. P. *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex* (Oxford Univ. Press, 1927).
2. Wolpe, J. *Psychotherapy by Reciprocal Inhibition* (Stanford Univ. Press, 1958).
3. Liu, X. *et al.* Optogenetic stimulation of a hippocampal engram activates fear memory recall. *Nature* **484,** 381–385 (2012).
4. Ramirez, S. *et al.* Creating a false memory in the hippocampus. *Science* **341,** 387–391 (2013).
5. Muramoto, K., Ono, T., Nishijo, H. & Fukuda, M. Rat amygdaloid neuron responses during auditory discrimination. *Neuroscience* **52,** 621–636 (1993).
6. Schoenbaum, G., Chiba, A. A. & Gallagher, M. Neural encoding in orbitofrontal cortex and basolateral amygdala during olfactory discrimination learning. *J. Neurosci.* **19,** 1876–1884 (1999).
7. Paton, J. J., Belova, M. A., Morrison, S. E. & Salzman, C. D. The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature* **439,** 865–870 (2006).
8. Shabel, S. J. & Janak, P. H. Substantial similarity in amygdala neuronal activity during conditioned appetitive and aversive emotional arousal. *Proc. Natl Acad. Sci. USA* **106,** 15031–15036 (2009).
9. Amano, T., Duvarci, S., Popa, D. & Paré, D. The fear circuit revisited: contributions of the basal amygdala nuclei to conditioned fear. *J. Neurosci.* **31,** 15481–15489 (2011).
10. Sangha, S., Chadick, J. Z. & Janak, P. H. Safety encoding in the basal amygdala. *J. Neurosci.* **33,** 3744–3751 (2013).
11. Teyler, T. J. & DiScenna, P. The hippocampal memory indexing theory. *Behav. Neurosci.* **100,** 147–154 (1986).
12. Rudy, J. W. & O'Reilly, R. C. Contextual fear conditioning, conjunctive representations, pattern completion, and the hippocampus. *Behav. Neurosci.* **113,** 867–880 (1999).
13. Reijmers, L. G., Perkins, B. L., Matsuo, N. & Mayford, M. Localization of a stable neural correlate of associative memory. *Science* **317,** 1230–1233 (2007).
14. Schiltz, C. A., Bremer, Q. Z., Landry, C. F. & Kelley, A. E. Food-associated cues alter forebrain functional connectivity as assessed with immediate early gene and proenkephalin expression. *BMC Biol.* **5,** 16 (2007).
15. Wheeler, A. L. *et al.* Identification of a functional connectome for long-term fear memory in mice. *PLOS Comput. Biol.* **9,** e1002853 (2013).
16. Kim, J., Kwon, J. T., Kim, H. S., Josselyn, S. A. & Han, J. H. Memory recall and modifications by activating neurons with elevated CREB. *Nature Neurosci.* **17,** 65–72 (2014).
17. Huff, M. L., Miller, R. L., Deisseroth, K., Moorman, D. E. & LaLumiere, R. T. Posttraining optogenetic manipulations of basolateral amygdala activity

modulate consolidation of inhibitory avoidance memory in rats. *Proc. Natl Acad. Sci. USA* **110,** 3597–3602 (2013).
18. Malkesman, O. *et al.* The female urine sniffing test: a novel approach for assessing reward-seeking behavior in rodents. *Biol. Psychiatry* **67,** 864–871 (2010).
19. Kiyokawa, Y., Hiroshima, S., Takeuchi, Y. & Mori, Y. Social buffering reduces male rats' behavioral and corticosterone responses to a conditioned stimulus. *Horm. Behav.* **65,** 114–118 (2014).
20. Maren, S. & Fanselow, M. S. Synaptic plasticity in the basolateral amygdala induced by hippocampal formation stimulation *in vivo*. *J. Neurosci.* **15,** 7548–7564 (1995).
21. Maren, S. & Hobin, J. A. Hippocampal regulation of context-dependent neuronal activity in the lateral amygdala. *Learn. Mem.* **14,** 318–324 (2007).
22. Herry, C. *et al.* Switching on and off fear by distinct neuronal circuits. *Nature* **454,** 600–606 (2008).
23. Stuber, G. D. *et al.* Excitatory transmission from the amygdala to nucleus accumbens facilitates reward seeking. *Nature* **475,** 377–380 (2011).
24. Han, J. H. *et al.* Neuronal competition and selection during memory formation. *Science* **316,** 457–460 (2007).
25. Choi, G. B. *et al.* Driving opposing behaviors with ensembles of piriform neurons. *Cell* **146,** 1004–1015 (2011).

Acknowledgements We thank X. Zhou, C. Potter, D. Plana, J. Martin, M. Tsitsiklis, H. Sullivan, W. Yu and A. Moffa for help with the experiments; K. L. Mulroy, T. Ryan and D. Roy for comments and discussions on the manuscript, and all the members of the Tonegawa laboratory for their support. This work was supported by the funds from the RIKEN Brain Science Institute, the Howard Hughes Medical Institute and The JPB Foundation to S.T., and the National Institutes of Health Pre-doctoral Training Grant T32GM007287 to J.K.

Author Contributions R.L.R., J.K. and S.T. contributed to the study design. R.L.R., J.K. and S.R. contributed to the data collection. X.L. cloned all constructs. R.L.R., J.K. and A.L.A. conducted the surgeries. R.L.R. and J.K. conducted the behavioural experiments. R.L.R. conducted the functional connectivity experiments. J.K. conducted the reversal experiments. R.L.R. contributed to the setup of the behavioural and optogenetic apparatus and programmed the behavioural software to run the experiments. R.L.R., J.K. and S.T. wrote the paper. All authors discussed and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.T. (tonegawa@mit.edu).

LETTER

doi:10.1038/nature13468

Exonuclease-mediated degradation of nascent RNA silences genes linked to severe malaria

Qingfeng Zhang[1,2,3], T. Nicolai Siegel[2,3]†, Rafael M. Martins[2,3], Fei Wang[1], Jun Cao[4], Qi Gao[4], Xiu Cheng[5], Lubin Jiang[5], Chung-Chau Hon[6], Christine Scheidig-Benatar[2,3], Hiroshi Sakamoto[2,3], Louise Turner[7], Anja T. R. Jensen[7], Aurelie Claes[2,3], Julien Guizetti[2,3], Nicholas A. Malmquist[2,3] & Artur Scherf[2,3]

Antigenic variation of the *Plasmodium falciparum* multicopy *var* gene family enables parasite evasion of immune destruction by host antibodies[1,2]. Expression of a particular *var* subgroup, termed *upsA*, is linked to the obstruction of blood vessels in the brain and to the pathogenesis of human cerebral malaria[3–6]. The mechanism determining *upsA* activation remains unknown. Here we show that an entirely new type of gene silencing mechanism involving an exonuclease-mediated degradation of nascent RNA controls the silencing of genes linked to severe malaria. We identify a novel chromatin-associated exoribonuclease, termed PfRNase II, that controls the silencing of *upsA var* genes by marking their transcription start site and intron-promoter regions leading to short-lived cryptic RNA. Parasites carrying a deficient *PfRNase II* gene produce full-length *upsA var* transcripts and intron-derived antisense long non-coding RNA. The presence of stable *upsA var* transcripts overcomes monoallelic expression, resulting in the simultaneous expression of both *upsA* and *upsC* type PfEMP1 proteins on the surface of individual infected red blood cells. In addition, we observe an inverse relationship between transcript levels of *PfRNase II* and *upsA*-type *var* genes in parasites from severe malaria patients, implying a crucial role of PfRNase II in severe malaria. Our results uncover a previously unknown type of post-transcriptional gene silencing mechanism in malaria parasites with repercussions for other organisms. Additionally, the identification of RNase II as a parasite protein controlling the expression of virulence genes involved in pathogenesis in patients with severe malaria may provide new strategies for reducing malaria mortality.

Beyond histone modifying enzymes[7–10], additional post-transcriptional mechanisms may control antigenic variation of the *P. falciparum* multicopy *var* gene family. RNA processing and degradation in *P. falciparum* erythrocytic-stage parasites has been inadequately studied, and therefore its potential role in the post-transcriptional control of virulence genes is unknown. We examined the RNA exosome, a RNase complex involved in RNA processing and decay associated with RNA quality control in the nucleus and cytoplasm of eukaryotic cells[11,12]. Bioinformatic analysis

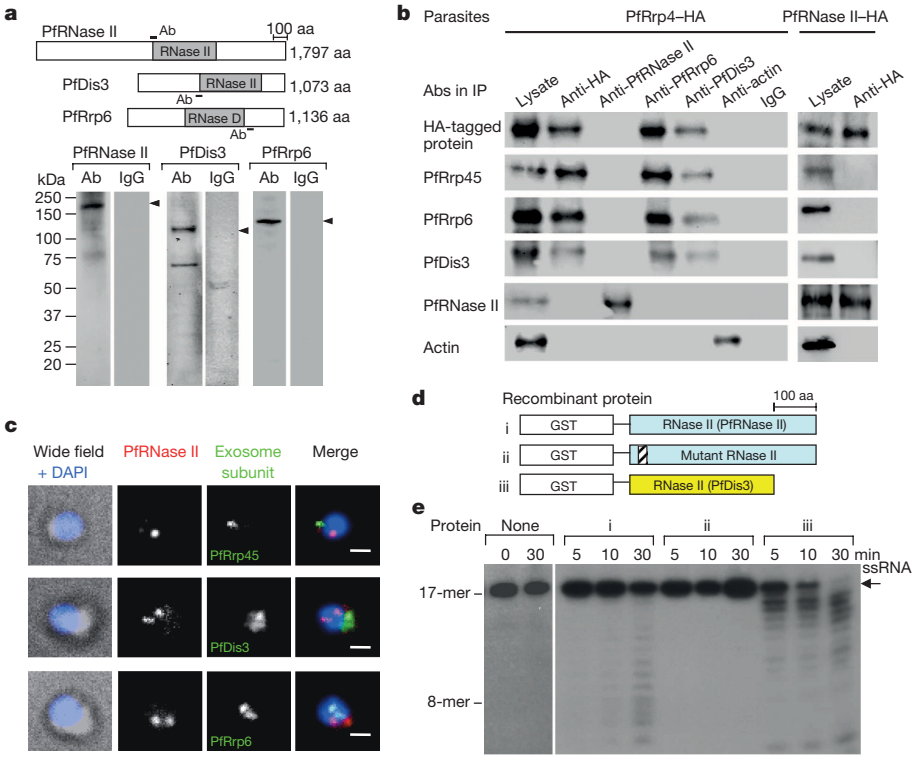


Figure 1 | Identification of a novel exoribonuclease (PfRNase II) in *P. falciparum*. **a**, Schematic representation and western blot analysis of three plasmodial exosome exoribonuclease-like proteins PfRNase II, PfDis3 and PfRrp6. Proteins corresponding to the predicted molecular masses are indicated by arrows. IgG, preimmune antibody control; aa, amino-acid residues; Ab, antibody. **b**, CoIP assay of PfRrp4–HA (core exosome subunit) and PfRNase II–HA transfectants with various antibodies as indicated for each immunoprecipitation (IP) reaction. The anti-PfActin I antibody was used as control. **c**, Co-immunofluorescence assay of three plasmodial exoribonucleases and a core exosome member, PfRrp45, in ring-stage wild-type 3D7 parasites. The rabbit anti-PfRrp45, anti-PfRNase II, anti-PfRrp6 and mouse anti-PfDis3 antibodies were used in these assays. DAPI, 4′,6-diamidino-2-phenylindole. Scale bar, 1 μm. **d**, Schematic diagram of recombinant RNase II domain of PfRNase II and PfDis3. Dead mutant of PfRNase II catalytic domain and GST were used as negative controls. **e**, Exoribonuclease activity analysis of recombinant RNase II domains *in vitro* with single-stranded RNA probe (ssRNA). Data in **a**–**c** and **e** are representative of two independent experiments.

[1]Research Center for Translational Medicine, Shanghai East Hospital and Institute of Infectious Diseases and Vaccine Development, Tongji University School of Medicine, Shanghai 200120, China. [2]Unité de Biologie des Interactions Hôte–Parasite, Institut Pasteur, F-75724 Paris, France. [3]CNRS, URA 2581, F-75724 Paris, France. [4]Jiangsu Institute of Parasitic Diseases, Key Laboratory of Parasitic Disease Control and Prevention (Ministry of Health), and Jiangsu Provincial Key Laboratory of Parasite Molecular Biology, Wuxi 214064, China. [5]Key Laboratory of Molecular Virology and Immunology, Institut Pasteur of Shanghai, Chinese Academy of Sciences, Shanghai 200031, China. [6]Cell Biology of Parasitism Unit, Institut Pasteur, and INSERM U786, F-75724 Paris, France. [7]Centre for Medical Parasitology, Department of International Health, Immunology & Microbiology, Faculty of Health and Medical Sciences, University of Copenhagen and Department of Infectious Diseases, Copenhagen University Hospital (Rigshospitalet), Copenhagen, Denmark. †Present address: Research Center for Infectious Diseases, University of Würzburg, 97080 Würzburg, Germany.

predicts eight putative RNA exosome-associated proteins in the *P. falciparum* genome, including exoribonuclease functional domain-containing proteins Dis3 and Rrp6 (Extended Data Table 1) (PlasmoDB site at http://plasmodb.org). Here we identify a non-canonical exoribonuclease containing a putative RNase II domain with no further homology to known exosome subunits that regulates *upsA var* gene expression, designated PfRNase II (PlasmoDB accession number PF3D7_0906000).

Structural prediction revealed homology between the RNase II domain of PfDis3 and human Dis3, and between PfRNase II and yeast Rrp44 (Extended Data Figs 1 and 2). By western blotting, antibodies against PfRNase II, PfDis3 and PfRrp6 reacted with bands of the predicted molecular masses 214, 135 and 126 kDa, respectively (Fig. 1a). To investigate the association between PfRNase II and core exosome complex, we generated a PfRrp4–haemagglutinin (HA)-tagged core exosome member transfectant and produced an antibody against another core exosome member, PfRrp45 (Extended Data Fig. 3a–d). Co-immunoprecipitation (CoIP) experiments showed that PfRNase II is a non-exosome exoribonuclease, whereas PfRrp6 and PfDis3 are the catalytic subunits of the plasmodial RNA exosome (Fig. 1b). Immunofluorescence assay (IFA) analysis revealed that PfRrp6 and PfRNase II localize as distinct foci at the nuclear periphery and that PfDis3 is localized in the cytoplasm adjacent to the nucleus at the ring stage (Fig. 1c). Immunoelectron microscopy confirmed the nuclear localization of PfRNase II (Extended Data Fig. 3e, f). To assess the RNA processing activity of the putative RNase II-like domain, we produced recombinant RNase II domains of PfRNase II and PfDis3 as glutathione *S*-transferase (GST) fusion proteins, using a PfRNase II-dead mutant and GST alone as negative controls (Fig. 1d and Extended Data Fig. 3g). *In vitro* RNA degradation assays revealed processive 3′–5′ hydrolytic activity towards single-stranded but not double-stranded RNA for PfRNase II and PfDis3, and no activity with the PfRNase II domain-dead mutant or GST alone (Fig. 1e and Extended Data Fig. 3h–j).

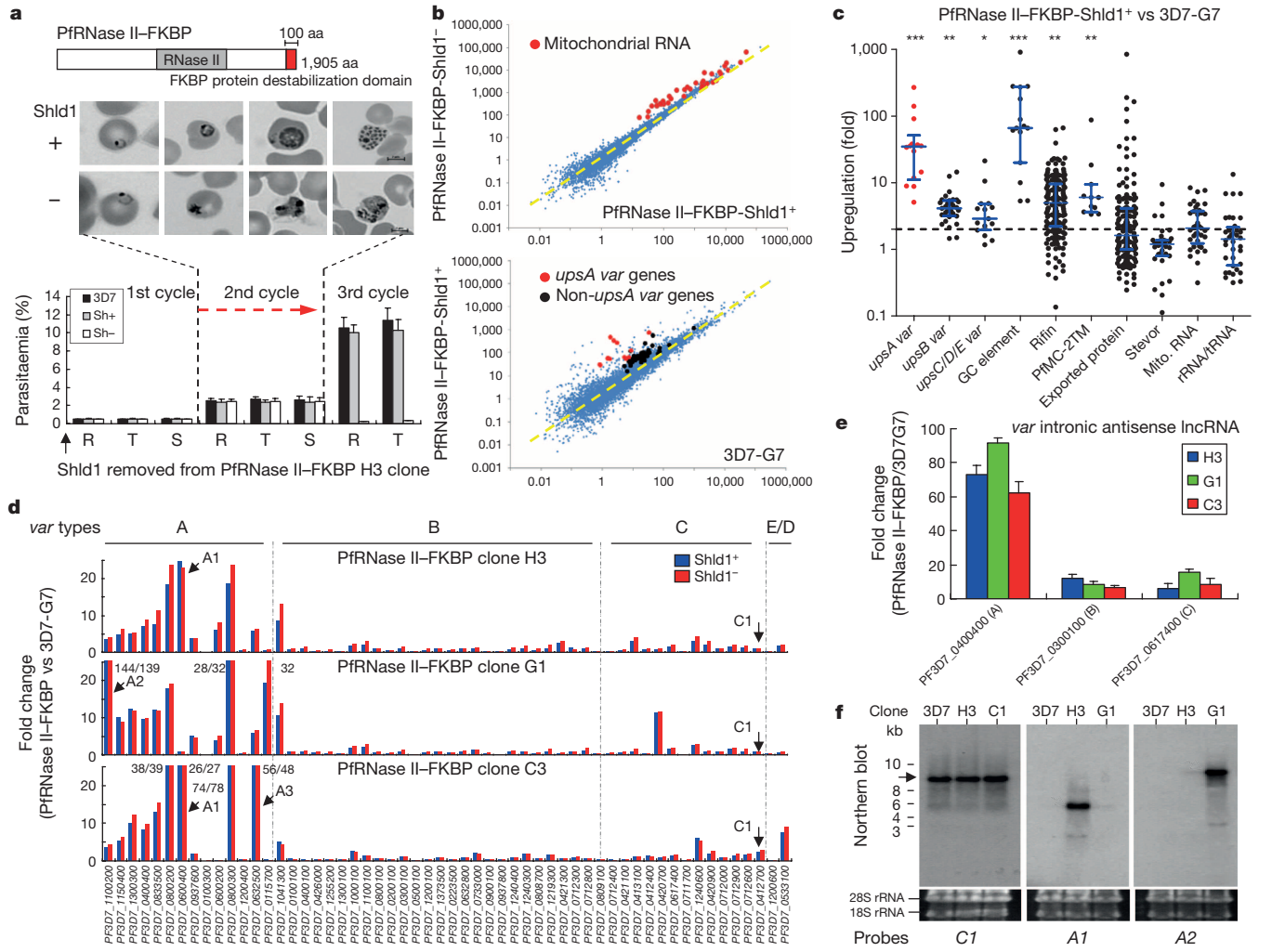


Figure 2 | Defect in PfRNase II leads to derepression of *upsA var* genes. **a**, Top: schematic diagram of PfRNase II–FKBP fusion protein. Bottom: growth curve of PfRNase II–FKBP clone H3 with or without Shld1. Parent 3D7-G7 was used as a control. Error bars represent s.e.m. for three independent experiments. **b**, Global transcriptome comparison of ring-stage parasite clones. Top: PfRNase II–FKBP Shld1^{+} versus Shld1^{-}. Bottom: PfRNase II–FKBP Shld1^{+} versus wild-type 3D7-G7. Both axes are logarithmic and correspond to normalized read numbers of individual genes shown as dots. **c**, Statistical analysis of transcript level changes in PfRNase II–FKBP Shld1^{+} over 3D7-G7. The medians are indicated as a line, and whiskers illustrate the interquartile range. The cutoff (twofold) is indicated by a dashed line. Mito., mitochondrial. ***, $P < 0.001$; **, $0.001 < P < 0.01$; *, $0.01 < P < 0.05$ (two-tailed Student's *t*-test). **d**, Transcriptional patterns of *var* gene family in ring-stage PfRNase II–FKBP-H3, G1, and C3 clones (Shld1^{+} or Shld1^{-}) determined by qPCR. The five *ups* subtypes are indicated on the top. The data are shown as fold change of PfRNase II–FKBP over 3D7-G7 related to the seryl-tRNA synthetase gene as an internal control. **e**, Transcriptional profiles of intronic antisense long non-coding RNA (lncRNA) from three individual *var* genes with different *ups*-types in PfRNase II–FKBP clones compared with 3D7-G7 measured by qPCR. Data are represented as means ± s.e.m. for three independent experiments. **f**, Northern blot assay of 3D7-G7 and PfRNase II–FKBP-H3 and G1 clones. The probe corresponding to each clone was *C1 var*, *A1 var* and *A2 var*, respectively. The full-length transcript of each *var* gene is indicated by an arrow. Loading control for total RNA is shown by ethidium bromide staining. Data in **a** and **f** are representative of three independent experiments.

Attempts to disrupt PfRNase II in parasites by using conventional knockout strategies failed, indicating that this gene is essential for parasite growth (Extended Data Fig. 4a). We therefore engineered a ligand-regulated FK506-binding protein (FKBP) destabilization domain (DD) fusion protein to investigate PfRNase II protein knockdown in a parasite clone (3D7-G7) expressing a single *upsC var* gene, Pf3D7_0412700) (Fig. 2a). The DD fusion protein is stabilized by the synthetic ligand Shld1 and is rapidly degraded in its absence[13]. Three independent PfRNase II–FKBP clones G1, H3 and C3 were obtained (Extended Data Fig. 4b, c), and showed comparable propagation rates during the first two erythrocytic cycles. Without Shld1, however, PfRNase II–FKBP protein levels were decreased and parasites revealed drastic morphological changes in mature stages of the second cycle and were deficient in progressing to a third cycle (Fig. 2a and Extended Data Fig. 4d). To study the transcriptional role of PfRNase II, we used RNA-seq analysis on total RNA from synchronized ring-stage parasites 48 h after the removal of Shld1 (Extended Data Fig. 5a). Transcription patterns for the PfRNase II–FKBP clone H3 in the presence or absence of Shld1 were comparable except for 22 out of 37 mitochondrial genes with transcriptional upregulation (at least twofold) (Fig. 2b and Extended Data Fig. 5b), possibly explaining the observed lethal phenotype. Next, we compared transcriptional profiles of PfRNase II–FKBP parasites with those of wild-type parasites (3D7-G7). We observed more than fivefold upregulation in transcript levels for 213 genes in PfRNase II–FKBP in comparison with wild-type parasites (Fig. 2b, Extended Data Fig. 5c and Supplementary Table 1). This suggests that the carboxy-terminal FKBP tagging of PfRNase II disrupts interactions that are important for gene silencing or that it interferes with RNase II activity. Almost all transcriptional changes were observed for clonally variant virulence genes and non-coding RNA transcribed from regions adjacent to *var* genes. Strikingly, 14 *upsA var* genes are upregulated 5-fold to 270-fold, whereas most other subtypes (*upsB* and *upsC*) showed relatively minor changes. Moreover, two non-coding RNAs encoded by the 15-member family of GC-rich elements adjacent to internal *var* genes[14] or from the subtelomeric non-coding TARE3 region showed strong upregulation in ring-stage parasites (Fig. 2c and Extended Data Fig. 5d). These data indicate the existence of a new type of post-transcriptional gene silencing of virulence genes.

To validate the observed transcriptional changes in severe malaria-associated *upsA var* genes, we performed real-time quantitative PCR (qPCR) for individual *var* genes in parasite clones. Despite an increase in *upsA var*-type transcript levels in PfRNase II–FKBP transgenic parasites, only few members reached transcript levels observed during monoallelic expression. Up to three distinct *upsA* members were found to be co-transcribed at very high levels together with the *upsC var C1* (3D7_0412700) expressed already in 3D7-G7 (Fig. 2d, f and Extended Data Fig. 6a). The *upsA* subgroup *var* genes are rarely activated in cultured parasites[15,16]. Each analysed clone expressed a different combination of *upsA* members, suggesting that switching occurs frequently in this subgroup, whereas *upsC var C1* expression remained stable. Control experiments (3D7 parasites grown with Shld1 or harbouring a non-integrated FKBP plasmid) revealed a single dominant *upsC* but no *upsA var* gene (Extended Data Fig. 6b). To further demonstrate dissociation of the

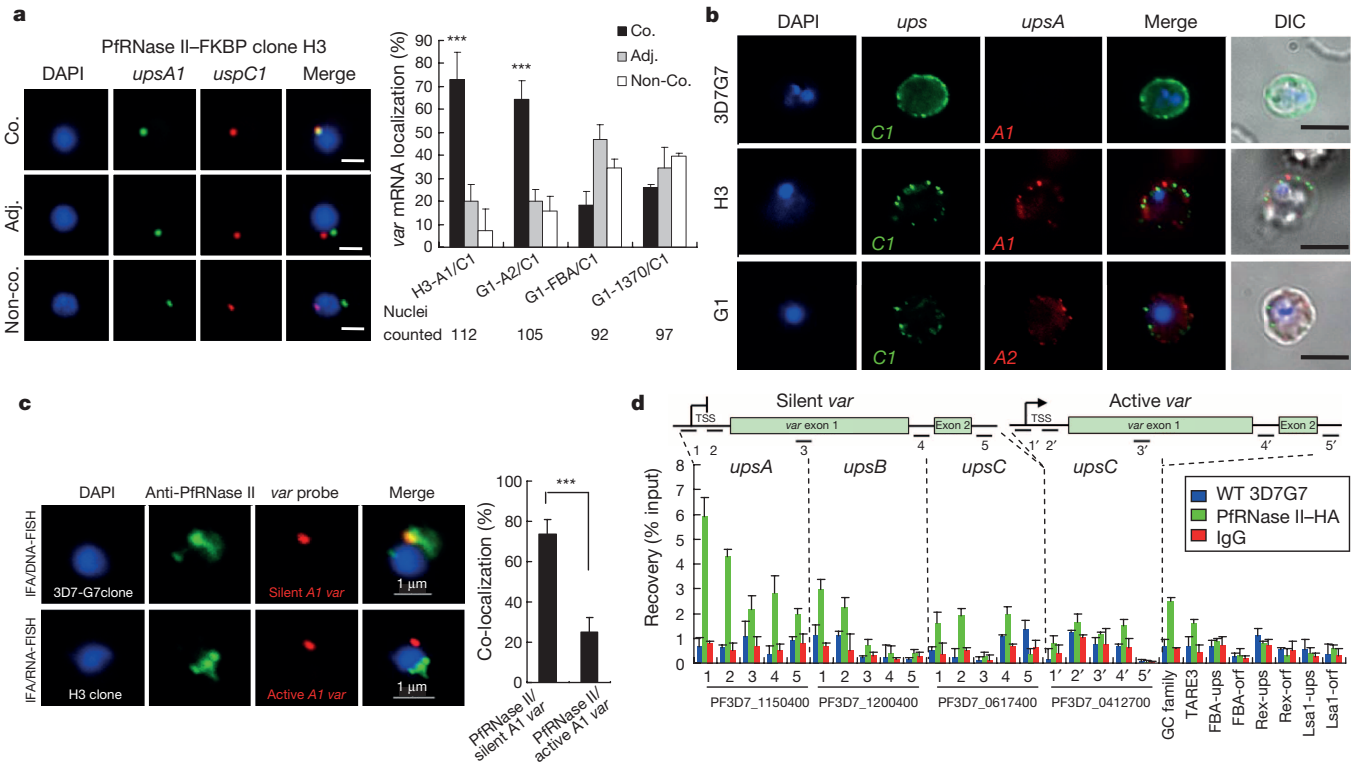


Figure 3 | Monoallelic expression of *var* gene family is controlled by chromatin-associated RNase II. **a**, Two-colour RNA-FISH (left) and statistical analysis of co-localization (right) of *var* transcripts in PfRNase II–FKBP-H3 and G1 clones. Red, *UpsC*-type *C1 var* probe; green, *upsA*-type *A1 var*, *A2 var* and control gene probes (fructose-bisphosphate aldolase (FBA), PF3D7_1444800; Glutaminyl-tRNA synthetase (1370): PF3D7_1331700). Co., co-localization; adj., adjacent; non-co., non-co-localization. ***, $P < 0.001$ (χ^2 test). Nuclear DNA was stained by DAPI (blue). Scale bar, 1 µm. Error bars represent s.e.m. for three independent experiments. **b**, Live-cell IFA using antisera to various PfEMP1s to detect co-expression of *var* genes in PfRNase II–FKBP H3 and G1 clones, respectively. The 3D7-G7 control expresses only one type of PfEMP1 (*C1 var*). Scale bar, 5 µm. **c**, PfRNase II is linked to the silent *var*-associated loci. Left: combined FISH/IFA assay using anti-PfRNase II antibody (IFA) with a FISH probe (PF3D7_0600400) that detected the *upsA var* expression site (RNA-FISH) in PfRNase II–FKBP clone H3. The same *upsA var* gene (DNA-FISH/IFA) was also analysed in ring-stage 3D7-G7 parasites. Right: statistical analysis of co-localization of each pair. In each analysis, more than 100 nuclei were counted; the error bars represent s.e.m. for three independent experiments. ***, $P < 0.001$ (χ^2 test). **d**, ChIP–qPCR of PfRNase II–HA transfectant. The enrichment of representative *var* genes with anti-HA antibody were shown respectively. The qPCR probes for each *var* gene are indicated by short lines under *var* gene loci. Error bars represent s.e.m. for three independent experiments. The data in **b** are representative of three independent experiments. TARE3, telomere-associated repeat element 3.

upsA var gene regulation, we cultured PfRNase II–FKBP parasite clones for a further 70 days (35 generations). qPCR analysis confirmed that most *upsA var* genes were upregulated in mutant clones but not in parent 3D7-G7 parasites, relative to day 0 (Extended Data Fig. 7a). For *upsB* and *upsC var* genes, some members showed higher transcript levels at day 70 but no striking differences between wild-type and mutant parasites. Subclone analysis of the C3 bulk culture (70 days) showed that new combinations of *upsA var* members co-transcribed with different members of *upsC* (Extended Data Fig. 6a). These data further illustrate the PfRNase II-dependent loss of monoallelic expression and continual switching of *upsA var* genes. In the PfRNase II–FKBP clones, a more than 60-fold transcriptional upregulation of antisense long non-coding RNA was observed from the *upsA var* gene intron region in ring-stage parasites, confirming the RNA sequencing result (Fig. 2e and Extended Data Fig. 8) and further highlighting a previously suggested role for exon 1 antisense long non-coding RNA in the process of *var* gene expression[7,17,18].

Single-cell analysis using two-colour RNA-fluorescence *in situ* hybridization (FISH) and surface immunofluorescence analysis confirmed the decoupling of *upsA var* genes from monoallelic expression in PfRNase II–FKBP parasites (Fig. 3a, b and Extended Data Fig. 7b). Using PfRNase II immunofluorescence with DNA-FISH (silent *var*) or RNA-FISH (active *var A1*), we observed a co-localization preference of PfRNase II protein to the silent *var* locus versus the active locus (Fig. 3c), indicating that removal of PfRNase II from the silent *upsA var* locus may trigger the gene activation *in situ*. We generated endogenous HA-tagged *PfRNase II* parasites to investigate its chromatin association by chromatin immunoprecipitation (ChIP)–qPCR, and confirmed that the C-terminal HA tag of PfRNase II did not interfere with monoallelic *var* expression (Extended Data Fig. 9a-d). ChIP analysis showed higher enrichment of PfRNase II protein at promoter and intron regions of *upsA var* genes, significantly less enrichment at other *var* gene types, and no association with the active one (Fig. 3d and Extended Data Fig. 9e). Given the potential role of PfRNase II in messenger RNA degradation, we performed nascent RNA analysis at *var* loci to study whether unstable mRNA was produced that was absent from steady-state RNA in wild-type 3D7-G7. We observed 15–20-fold higher levels of 5′ *upsA var* transcripts in nascent RNA than in total RNA. Non-*upsA var* genes and control genes showed slight, if any, difference between nascent and steady-state RNA levels (Fig. 4a), indicating cryptic mRNA production from *upsA var* genes. Further analysis revealed that both upstream region and intron promoter-derived antisense long non-coding RNA were the main sources of cryptic RNAs produced by *upsA var* genes (Fig. 4b and Extended Data Fig. 7c).

To investigate the clinical relevance of our findings, we extracted total *P. falciparum* RNA from blood from patients with severe malaria and from patients with uncomplicated malaria. As observed previously[19], *upsA* transcript levels were higher in patients with severe malaria than in those with uncomplicated malaria (Extended Data Fig. 7d), but *PfRNase II* transcripts were inversely correlated with *upsA* transcript levels (Fig. 4c), thereby linking decreased PfRNase II expression to *upsA var* gene regulation in patients with severe malaria.

Our findings identify a novel epigenetic silencing pathway in malaria parasites that is distinct from the post-transcriptional repression mechanism described previously[20]. Silenced *upsA var* genes continually produce short-lived sense and antisense transcripts from promoter and intron regions as a result of rapid *in situ* degradation by a novel type of RNase II. This may explain the several *var* transcripts detected in single infected erythrocytes (ring stage) by PCR[21]. By producing RNA from virtually all *upsA var* genes in RNase II-deficient parasites, singular expression of *var* genes is abolished. This raises the possibility that antisense long non-coding RNA is involved in the *var* promoter activation process. Our data show a resemblance to the degradation of cryptic unstable nuclear RNA by the exosome observed in yeast[22]. However, PfRNase II exerts its nuclear function dissociated from the RNA exosome core at specific chromatin regions (Extended Data Fig. 9f). We assume that the selective removal of PfRNase II at a specific *upsA*

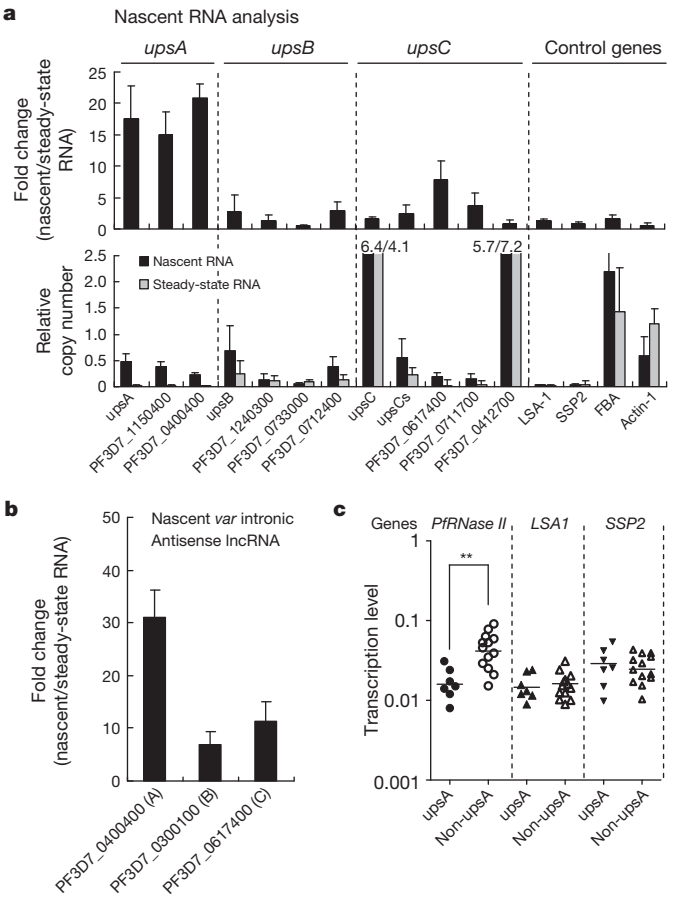


Figure 4 | *In situ* degradation of nascent RNA at promoter regions leads to silencing of severe malaria-associated *upsA*-type *var* genes. **a**, Comparative qPCR *var* transcription analysis from nascent and steady-state mRNA in ring-stage 3D7-G7 wild-type parasites. Primers are designed against the upstream promoter regions of *ups*-specific or individual *var* and control genes. *upsCs* primers are silent *upsC var*-specific. The seryl-tRNA synthetase gene was used as an internal control for the calculation of relative copy numbers. Error bars represent s.e.m. for three independent experiments. **b**, Transcription profiles of nascent intronic antisense long non-coding RNA (lncRNA) from three individual *var* genes with different *ups* types from the steady-state RNA measured by qPCR assay. Transcription levels are shown as fold change of nascent RNA over total RNA. Error bars represent s.e.m. for three independent experiments. **c**, Transcription analysis of *PfRNase II* gene in field isolates with higher transcript levels of *upsA var* genes (from patients with severe malaria) and with lower levels (from patients with uncomplicated malaria). The *LSA-1* and *SSP2* genes with comparable transcriptional levels to those of *PfRNase II* are used as controls. **, $P < 0.01$ (two-tailed Student's *t*-test).

var locus allows expression of that locus, a notion supported by our co-localization experiments (Fig. 3c). Upregulation of other types of RNA in PfRNase II mutant parasites indicates that this type of mechanism might be a wide-ranging gene expression control mechanism.

In eukaryotic organisms, monoallelic expression controls distinct biological processes. For example in mice, only one of the ~1,400 genes encoding olfactory receptors is expressed in any given olfactory sensory neuron[23]. Similarly, protozoan parasites that cause malaria, giardiasis or sleeping sickness express one of many possible variant surface antigen genes at a time[24]. This work identifies an entirely new mechanism contributing to monoallelic gene expression in malaria parasites that may be relevant in other organisms. In addition, the identification of the first *Plasmodium* protein controlling virulence factors expressed in patients with severe malaria may provide new strategies for reducing malaria pathogenesis.

METHODS SUMMARY

The association of exosome-like exoribonucleases and the core exosome complex was investigated by CoIP assays. FKBP destabilization domain and triple HA tagging at the C terminus of *PfRNase II* gene were carried out using the single-crossover recombinant strategy. The PfRNase II–FKBP clones were used in the comparative transcriptome analysis with wild-type 3D7-G7 clone by single-stranded RNA sequencing analysis[25]. The phenotypes of PfRNase II–FKBP clones were determined by RNA-FISH, IFA and fluorescence-activated cell sorting (FACS) with individual PfEMP1-specific antibodies in live infected red blood cells (iRBCs). The chromatin-associated PfRNase II was validated by electron microscopy, RNA-FISH, DNA-FISH and ChIP–qPCR analysis. We isolated the nascent RNA in a 3D7-G7 clone and used qPCR to measure the level of various regions of *upsA var* genes with other type *var* genes and housekeeping genes as controls, and compared them with those of steady-state RNA. All the primers used in this study are shown in Supplementary Table 2.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 October 2013; accepted 12 May 2014.
Published online 29 June 2014.

1. Miller, L. H., Baruch, D. I., Marsh, K. & Doumbo, O. K. The pathogenic basis of malaria. *Nature* **415,** 673–679 (2002).
2. Scherf, A., Riviere, L. & Lopez-Rubio, J. J. SnapShot: *var* gene expression in the malaria parasite. *Cell* **134,** 190–190.e1 (2008).
3. Deitsch, K. W. & Chitnis, C. E. Molecular basis of severe malaria. *Proc. Natl Acad. Sci. USA* **109,** 10130–10131 (2012).
4. Avril, M. *et al.* A restricted subset of *var* genes mediates adherence of *Plasmodium falciparum*-infected erythrocytes to brain endothelial cells. *Proc. Natl Acad. Sci. USA* **109,** E1782–E1790 (2012).
5. Claessens, A. *et al.* A subset of group A-like *var* genes encodes the malaria parasite ligands for binding to human brain endothelial cells. *Proc. Natl Acad. Sci. USA* **109,** E1772–E1781 (2012).
6. Turner, L. *et al.* Severe malaria is associated with parasite binding to endothelial protein C receptor. *Nature* **498,** 502–505 (2013).
7. Jiang, L. *et al.* PfSETvs methylation of histone H3K36 represses virulence genes in *Plasmodium falciparum*. *Nature* **499,** 223–227 (2013).
8. Lopez-Rubio, J. J., Mancio-Silva, L. & Scherf, A. Genome-wide analysis of heterochromatin associates clonally variant gene regulation with perinuclear repressive centers in malaria parasites. *Cell Host Microbe* **5,** 179–190 (2009).
9. Freitas-Junior, L. H. *et al.* Telomeric heterochromatin propagation and histone acetylation control mutually exclusive expression of antigenic variation genes in malaria parasites. *Cell* **121,** 25–36 (2005).
10. Duraisingh, M. T. *et al.* Heterochromatin silencing and locus repositioning linked to regulation of virulence genes in *Plasmodium falciparum*. *Cell* **121,** 13–24 (2005).
11. Jacquier, A. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nature Rev. Genet.* **10,** 833–844 (2009).
12. Kiss, D. L. & Andrulis, E. D. The exozyme model: a continuum of functionally distinct complexes. *RNA* **17,** 1–13 (2011).
13. Armstrong, C. M. & Goldberg, D. E. An FKBP destabilization domain modulates protein levels in *Plasmodium falciparum*. *Nature Methods* **4,** 1007–1009 (2007).
14. Mourier, T. *et al.* Genome-wide discovery and verification of novel structured RNAs in *Plasmodium falciparum*. *Genome Res.* **18,** 281–292 (2008).
15. Frank, M., Dzikowski, R., Amulic, B. & Deitsch, K. Variable switching rates of malaria virulence genes are associated with chromosomal position. *Mol. Microbiol.* **64,** 1486–1498 (2007).
16. Horrocks, P., Pinches, R., Christodoulou, Z., Kyes, S. A. & Newbold, C. I. Variable *var* transition rates underlie antigenic variation in malaria. *Proc. Natl Acad. Sci. USA* **101,** 11129–11134 (2004).
17. Ralph, S. A. *et al.* Transcriptome analysis of antigenic variation in *Plasmodium falciparum-var* silencing is not dependent on antisense RNA. *Genome Biol.* **6,** R93 (2005).
18. Epp, C., Li, F., Howitt, C. A., Chookajorn, T. & Deitsch, K. W. Chromatin associated sense and antisense noncoding RNAs are transcribed from the *var* gene family of virulence genes of the malaria parasite *Plasmodium falciparum*. *RNA* **15,** 116–127 (2009).
19. Jensen, A. T. *et al. Plasmodium falciparum* associated with severe childhood malaria preferentially expresses PfEMP1 encoded by group A *var* genes. *J. Exp. Med.* **199,** 1179–1190 (2004).
20. Mair, G. R. *et al.* Regulation of sexual development of *Plasmodium* by translational repression. *Science* **313,** 667–669 (2006).
21. Chen, Q. *et al.* Developmental selection of var gene expression in *Plasmodium falciparum*. *Nature* **394,** 392–395 (1998).
22. Gudipati, R. K., Neil, H., Feuerbach, F., Malabat, C. & Jacquier, A. The yeast RPL9B gene is regulated by modulation between two modes of transcription termination. *EMBO J.* **31,** 2427–2437 (2012).
23. Chess, A., Simon, I., Cedar, H. & Axel, R. Allelic inactivation regulates olfactory receptor gene expression. *Cell* **78,** 823–834 (1994).
24. Deitsch, K. W., Lukehart, S. A. & Stringer, J. R. Common strategies for antigenic variation by bacterial, fungal and protozoan pathogens. *Nature Rev. Microbiol.* **7,** 493–503 (2009).
25. Siegel, T. N. *et al.* Strand-specific RNA-Seq reveals widespread and developmentally regulated transcription of natural antisense transcripts in *Plasmodium falciparum*. *BMC Genomics* **15,** 150 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank T. Wandless for providing Shld1 compound. This work was supported by a European Research Council Advanced Grant (PlasmoEscape 250320), the French Parasitology consortium ParaFrap (ANR-11-LABX0024), the National Natural Science Foundation of China (NSFC; no. 31271388), the French National Research Agency (13-ISV3-0003-01)–NSFC (no. 81361130411) International Collaboration Project, and the Fundamental Research Funds for the Central Universities of China (20123283). T.N.S. was supported by the Human Frontier Science Program and a European Molecular Biology Organization long-term fellowship. J.C. was supported by the NSFC (no. 81271870). J.G. was supported by the Human Frontier Science Program.

Author Contributions Q.Z. and A.S. conceived and designed experiments. Q.Z., T.N.S. and R.M.M. performed most of the experiments. Q.Z., T.N.S. and C.H. performed RNA sequencing and data analysis. R.M.M. produced recombinant proteins and performed the exoribonuclease assay *in vitro*. L.J., X.C., F.W. and H.S. generated constructions, transfectants and parasite material. J.C. and Q.G. collected the field isolates and performed gene transcription analysis. C.S. performed the northern blot assay. L.T. and A.T.R.J. generated the antibodies against individual PfEMP1. J.G. and N.A.M. performed IFA and FACS. Q.Z. and A.S. analysed all the data and wrote the manuscript. All authors discussed and approved the manuscript.

Author Information The RNA-seq data generated for this study have been deposited in the European Nucleotide Archive under accession number PRJEB4511, and RNA-seq data from wild-type 3D7 cells[25] used for this study are available under accession no. ERP001849. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.S. (artur.scherf@pasteur.fr) or Q.Z. (qfzhangsh@aliyun.com).

LETTER

doi:10.1038/nature13540

Endocrinization of FGF1 produces a neomorphic and potent insulin sensitizer

Jae Myoung Suh[1]*, Johan W. Jonker[2]*, Maryam Ahmadian[1], Regina Goetz[3], Denise Lackey[4], Olivia Osborn[4], Zhifeng Huang[3]†, Weilin Liu[2], Eiji Yoshihara[1], Theo H. van Dijk[2], Rick Havinga[2], Weiwei Fan[1], Yun-Qiang Yin[1], Ruth T. Yu[1], Christopher Liddle[5], Annette R. Atkins[1], Jerrold M. Olefsky[4], Moosa Mohammadi[3], Michael Downes[1] & Ronald M. Evans[1,6]

Fibroblast growth factor 1 (FGF1) is an autocrine/paracrine regulator whose binding to heparan sulphate proteoglycans effectively precludes its circulation[1,2]. Although FGF1 is known as a mitogenic factor, FGF1 knockout mice develop insulin resistance when stressed by a high-fat diet, suggesting a potential role in nutrient homeostasis[3,4]. Here we show that parenteral delivery of a single dose of recombinant FGF1 (rFGF1) results in potent, insulin-dependent lowering of glucose levels in diabetic mice that is dose-dependent but does not lead to hypoglycaemia. Chronic pharmacological treatment with rFGF1 increases insulin-dependent glucose uptake in skeletal muscle and suppresses the hepatic production of glucose to achieve whole-body insulin sensitization. The sustained glucose lowering and insulin sensitization attributed to rFGF1 are not accompanied by the side effects of weight gain, liver steatosis and bone loss associated with current insulin-sensitizing therapies. We also show that the glucose-lowering activity of FGF1 can be dissociated from its mitogenic activity and is mediated predominantly via FGF receptor 1 signalling. Thus we have uncovered an unexpected, neomorphic insulin-sensitizing action for exogenous non-mitogenic human FGF1 with therapeutic potential for the treatment of insulin resistance and type 2 diabetes.

Increases in the prevalence of obesity and insulin resistance and the associated incidence of difficult-to-manage type 2 diabetes have become world-wide public health problems as well as a financial burden for the health care system, emphasizing the urgent need for improved insulin-sensitizing therapies. Thiazolidinediones are highly effective oral medications for type 2 diabetes that act through the nuclear receptor peroxisome proliferator activated receptor γ (PPARγ) to control networks of genes involved in adipogenesis, lipid metabolism and insulin sensitization. However, the unique sensitization benefits of thiazolidinediones are compromised by detrimental side effects, including weight gain, bone loss and congestive heart failure, suggesting that targeting downstream effectors of PPARγ may evoke fewer side effects yet retain insulin-sensitizing potential[5]. In this regard, FGF21, a member of the endocrine FGF subfamily whose expression is regulated by PPARγ in adipose tissue, has been identified as an effective glucose-lowering agent in rodents[6,7]. Recently, FGF1, the prototype of the FGF family of proteins, was also found to be transcriptionally regulated by PPARγ in adipose tissue, and *Fgf1* knockout mice exhibit an aggressive insulin-resistant phenotype when stressed by a high-fat diet[3,4]. The affinity of FGF1 for heparan sulphate proteoglycans results in autocrine/paracrine signalling and limited serum exposure, distinguishing it from the endocrine FGFs, which include FGF21 (refs 1, 2, 8, 9). These studies, combined with those demonstrating that FGF receptor agonist-antibodies modulate glucose homeostasis[10] raised the question as to whether endocrinization of the non-endocrine FGF1 could elicit glucose-lowering effects.

To explore its therapeutic potential, recombinant murine FGF1 (rFGF1) was injected subcutaneously (subQ) into genetically induced (*ob*/*ob* and *db*/*db*) as well as diet-induced obese (DIO) insulin-resistant mice. We found that a single injection of 0.5 mg kg^{-1} rFGF1 was sufficient to attain normoglycaemia in the severely hyperglycaemic *ob*/*ob* mice (Fig. 1a). Maximal glucose lowering was achieved within 18–24 h, and sustained effects observed for more than 48 h. Moreover, this effect was dose dependent, but even at the maximal dose (2.0 mg kg^{-1}) it did not result in hypoglycaemia (Fig. 1b and data not shown). Potent glucose lowering was

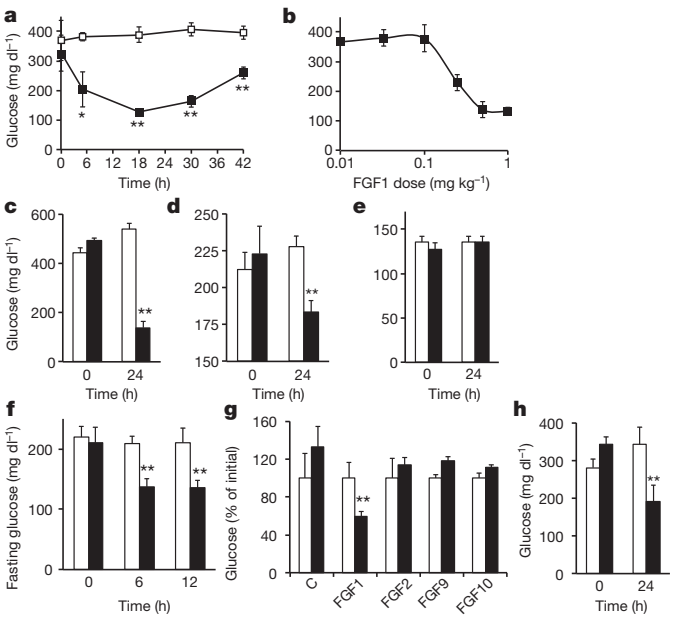


Figure 1 | Acute rFGF1 injection lowers glucose in diabetic mice. **a**, Blood glucose levels in *ob*/*ob* mice after rFGF1 (filled squares, $n = 6$) or control vehicle (open squares, $n = 4$) injection. **b**, Dose response of rFGF1 on blood glucose levels in *ob*/*ob* mice 24 h after injection ($n = 3$). **c**–**f**, Blood glucose levels after rFGF1 injection in *db*/*db* ($n = 3$) (**c**), DIO ($n = 6$) (**d**), normal-chow-fed ($n = 8$) (**e**) and fasted *ob*/*ob* ($n = 6$ for control; $n = 5$ for rFGF1) (**f**) mice. **g**, Blood glucose levels in *ob*/*ob* mice before (open bars) and 24 h after (filled bars) injection of murine FGF peptides (control, FGF1, FGF2: $n = 4$; FGF9, FGF10: $n = 2$). **h**, Blood glucose levels in *ob*/*ob* mice after human rFGF1 (filled bars, $n = 4$) or control vehicle (open bars, $n = 4$). C, control. Recombinant FGF peptides (0.5 mg kg^{-1}) or control vehicle (PBS) were injected subcutaneously to *ad libitum* fed mice unless otherwise noted. Values are means and s.e.m. Statistics by two-tailed *t*-test: $^*P < 0.05$; $^{**}P < 0.01$.

[1]Gene Expression Laboratory, Salk Institute for Biological Studies, La Jolla, California 92037, USA. [2]Center for Liver, Digestive and Metabolic Diseases, Departments of Pediatrics and Laboratory Medicine, University of Groningen, University Medical Center Groningen, Hanzeplein 1, 9713 GZ Groningen, The Netherlands. [3]Department of Biochemistry and Molecular Pharmacology, New York University School of Medicine, New York, New York 10016, USA. [4]Department of Medicine, Division of Endocrinology and Metabolism, University of California at San Diego, La Jolla, California 92093, USA. [5]The Storr Liver Unit, Westmead Millennium Institute and University of Sydney, Westmead Hospital, Westmead, New South Wales 2145, Australia. [6]Howard Hughes Medical Institute, Salk Institute for Biological Studies, La Jolla, California 92037, USA. †Present address: School of Pharmacy, Wenzhou Medical University, Wenzhou, Zhejiang 325035, China.
*These authors contributed equally to this work.

observed in both *db*/*db* and DIO mouse models, and rFGF1 was effective when delivered either intraperitoneally or intravenously in *ob*/*ob* mice, independently of exogenous heparin (Fig. 1c, d and Extended Data Fig. 1a–c). rFGF1 had no effect on blood glucose or insulin levels in normoglycaemic chow-fed mice (Fig. 1e and Extended Data Fig. 1d, e). Consistent with the known effects of FGF1 on feeding[11–13], we observed a transient suppression of food intake that correlated with a temporary decrease in body weight (Extended Data Fig. 1f–i and data not shown). However, injection of rFGF1 similarly reduced glucose levels under fasting conditions, dissociating the glucose-lowering effects of rFGF1 from its effect on food intake (Fig. 1f). FGF1 is considered the universal ligand for FGF receptors (FGFRs) in its ability to bind and activate each of the alternatively spliced forms of the four tyrosine kinase FGF receptors (FGFR1–FGFR4), whereas other members of the FGF superfamily demonstrate receptor specificity[1]. To determine whether other autocrine/paracrine FGFs have similar blood glucose lowering activity when given pharmacologically, we tested a selection of FGFs with specificities covering all seven FGFR receptors. FGF1 seems unique in its ability to lower blood glucose: FGF2, FGF9 and FGF10 failed to do so (Fig. 1g). Furthermore, recombinant human FGF1 (hFGF1) was similarly able to normalize blood glucose in diabetic mice, suggesting an evolutionarily conserved pathway (Fig. 1h).

The pronounced glucose-lowering efficacy of a single injection of rFGF1 led us to investigate the effects of serial injections. *ob*/*ob* mice injected with 0.5 mg kg^{-1} every other day for 35 days (chronic treatment) did not develop any apparent resistance to rFGF1, exhibiting sustained glucose lowering with minimal changes in body weight or composition (Fig. 2a, b and Extended Data Fig. 2a, b). A similar glucose-lowering effect was also seen in pair-fed *ob*/*ob* mice, indicating that the transient reduction in food intake after chronic treatment with rFGF1 does not account for the beneficial glucose-lowering effects (Extended Data Fig. 2c, d). The fasting blood glucose levels of chronically rFGF1-treated *ob*/*ob* mice were 50% lower than in PBS-treated control mice, and remained lower throughout the glucose tolerance test (GTT) with a coincident decrease in insulin levels. Furthermore, rFGF1-treated mice showed a marked improvement in insulin sensitivity as measured by an insulin tolerance test (ITT) (Fig. 2c–e). Although there was no significant effect on free fatty acids, cholesterol and triglyceride levels in serum (Extended Data Table 1 and Extended Data Fig. 2e, f), chronic treatment with rFGF1 decreased hepatic steatosis and increased liver glycogen content, as shown histologically and by quantitative analyses (Fig. 2f–h). No significant changes were detected in serum metabolic hormone levels (Extended Data Fig. 2g). Furthermore, chronic treatment with rFGF1 of DIO mice, a strain that more closely models the majority of human type 2 diabetes, also resulted in pronounced and sustained lowering of blood glucose levels (Fig. 2i) as well as in increased insulin sensitization as measured by GTT and ITT (Fig. 2j, k). These beneficial effects in DIO mice were observed without significant changes in body weight, organ weights and feeding trends (Extended Data Fig. 2h–j).

The potent and rapid normalization of blood glucose by parenteral rFGF1 led us to investigate potential insulin secretagogue or insulin-mimetic activities of rFGF1. rFGF1 administered by injection had no significant effect on glucose-stimulated insulin secretion in isolated pancreatic islets and did not increase serum insulin levels under basal conditions or during a GTT, indicating that exogenous FGF1 does not stimulate pancreatic β-cell insulin release, either *ex vivo* or *in vivo* (Fig. 2d and Extended Data Fig. 3a–c). Next, we tested rFGF1 efficacy in mice rendered diabetic by streptozotocin-induced destruction of insulin-producing β-cells. In this diabetic mouse model (STZ mice), rFGF1 alone failed to lower blood glucose levels, indicating that rFGF1 is not an insulin mimetic (Fig. 3a). However, pretreatment of STZ mice with rFGF1 markedly enhanced the glucose-lowering effects of exogenously supplied insulin (Fig. 3b, c). Conversely, the rFGF1-dependent increase in metabolic clearance rate was severely blunted when insulin secretion was inhibited by somatostatin (Fig. 3d–g). Improved sensitivity to insulin often correlates with reduced systemic inflammation; indeed, chronic treatment of *ob*/*ob*

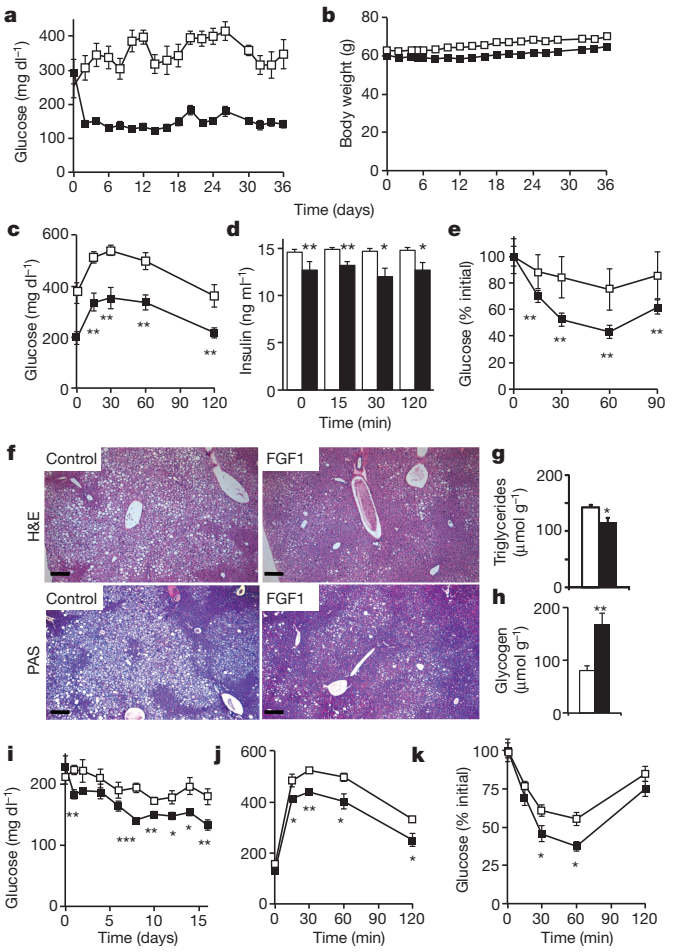


Figure 2 | Chronic administration of rFGF1 achieves sustained glucose lowering and insulin sensitization in diabetic mice. **a**, **b**, Random-fed blood glucose (**a**) and body weight (**b**) during chronic treatment of *ob*/*ob* mice with rFGF1. **c**–**e**, GTT (**c**), insulin levels during GTT (**d**) and ITT (**e**) measured in *ob*/*ob* mice after 4 weeks of rFGF1 treatment (control $n = 6$, rFGF1 $n = 8$). **f**–**h**, Representative haematoxylin and eosin (H&E) and periodic acid–Schiff staining (PAS; magenta represents glycogen) (**f**), and triglyceride (**g**) and glycogen (**h**) content of *ob*/*ob* livers after 5 weeks of chronic treatment with rFGF1 ($n = 6$ for overnight fasted control; $n = 8$ for rFGF1). **i**–**k**, Random-fed blood glucose (**i**), GTT (**j**) and ITT (**k**) in DIO mice after 3 weeks rFGF1 treatment ($n = 6$). *Ad libitum* fed mice were injected subcutaneously with 0.5 mg kg^{-1} rFGF1 (filled bars and symbols) or control vehicle (PBS, open bars and symbols) every 48 h. Values are means and s.e.m. Statistics by two-tailed *t*-test: $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.005$. Scale bar, 200 μm.

mice with rFGF1 decreased serum levels of several inflammatory cytokines (eotaxin, keratinocyte chemoattractant (KC), Mip-1b and interleukin-3) (Extended Data Fig. 3d). These results demonstrate that rFGF1 operates in an insulin-dependent manner to lower blood glucose, and suggest that parenteral rFGF1 may act as an insulin sensitizer.

To investigate the physiological mechanisms of rFGF1 action further, we performed hyperinsulinaemic–euglycaemic clamp studies in DIO mice chronically treated with vehicle or rFGF1. The steady-state glucose infusion rate during the clamp was ~75% higher in rFGF1-injected mice, indicating increased responsiveness to insulin (Extended Data Fig. 4a). The ability of insulin to suppress hepatic glucose production was improved in rFGF1-injected mice, revealing increased hepatic insulin sensitivity as a long-term consequence of rFGF1 injection (Fig. 4a). Liver gene expression analyses from chronically rFGF1-treated DIO mice revealed significant reductions in macrophage markers, for example *F4/80* and *Cd11c*, and in inflammatory cytokines, for example *Il-1α*, *Il-1β* and *Tnf-α* (Extended Data Fig. 4b). Furthermore, whole-body and insulin-stimulated glucose disposal rates were ~47% and ~80% higher, respectively, in

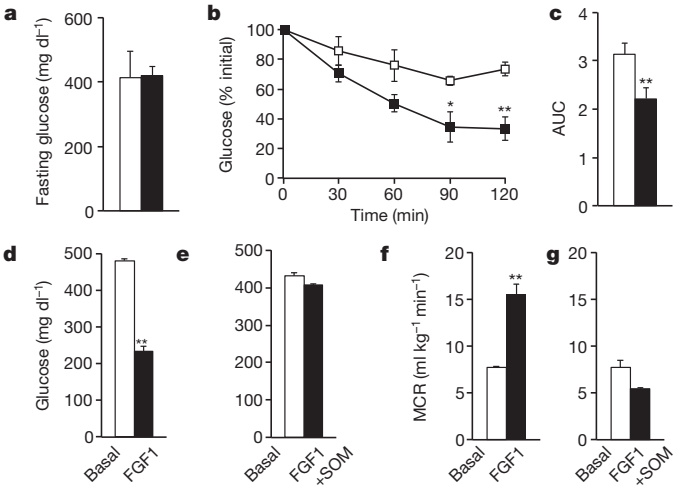


Figure 3 | rFGF1 induces insulin-dependent glucose uptake. **a**, Blood glucose levels in STZ-induced diabetic mice 8 h after subcutaneous injection of control PBS (open bars) or rFGF1 (0.5 mg kg^{-1}, filled bars). **b**, **c**, ITT (**b**) and area under the ITT curve (AUC; **c**) for control and rFGF1-treated STZ mice (open and filled squares and bars, respectively; $n = 4$). **d**–**g**, Blood glucose level (**d**, **e**) and metabolic clearance rate (MCR; **f**, **g**) in *ob*/*ob* mice after control PBS (open bars) or rFGF1 injection (0.2 mg kg^{-1} intravenously, filled bars) before (**d**, **f**) or 1 h after (**e**, **g**) infusion of somatostatin (SOM) ($n = 5$). Values are means and s.d. Statistics by two-tailed *t*-test: $*P < 0.05$; $**P < 0.01$.

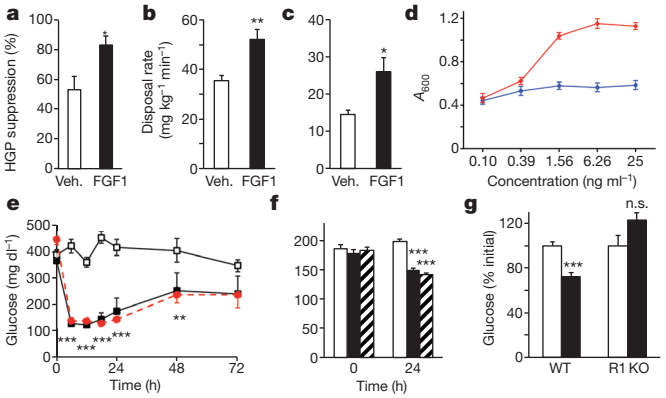


Figure 4 | Chronic administration of rFGF1 is insulin sensitizing. **a**–**c**, Insulin-stimulated suppression of hepatic glucose production (HGP) (**a**), steady-state glucose disposal rate (**b**) and insulin-stimulated glucose disposal rate (**c**) measured during hyperinsulinaemic–euglycaemic clamps on DIO mice after 3 weeks of control PBS (Veh., open bars, $n = 11$) or rFGF1 (0.5 mg kg^{-1} subcutaneously every other day; filled bars, $n = 9$) treatment. **d**, Proliferative activity (measured by the absorbance of formazan at 600 nm) of NIH3T3 cells treated with rFGF1 (red) or rFGF1$^{\Delta NT}$ (blue) at the indicated concentrations (experiment repeated three times). **e**, Blood glucose levels of *ob*/*ob* mice treated with control PBS (open symbols, $n = 12$), rFGF1 (0.5 mg kg^{-1} subcutaneously; filled symbols, $n = 8$) or rFGF1$^{\Delta NT}$ (0.5 mg kg^{-1} subcutaneously; red dashed line, $n = 6$). **f**, Blood glucose levels of DIO mice treated with control PBS (open bars), rFGF1 (0.5 mg kg^{-1} subcutaneously, filled bars), or rFGF1$^{\Delta NT}$ (0.5 mg kg^{-1} subcutaneously, striped bars) at the indicated times ($n = 10$). **g**, Blood glucose levels in 8-month-old mice fed on a high-fat diet: *Fgfr1*$^{fl/fl}$ (WT, $n = 5$) and aP2-Cre;*Fgfr1*$^{fl/fl}$ (R1 knockout (KO), $n = 4$) mice at 0 h (open bars) and 24 h (filled bars) after treatment with rFGF1 (0.5 mg kg^{-1} subcutaneously). Values are means and s.e.m. Statistics by two-tailed *t*-test: $*P < 0.05$; $**P < 0.01$; $***P < 0.005$; n.s., not significant.

rFGF1-injected mice, reflecting enhanced peripheral sensitivity to insulin (Fig. 4b, c and Extended Data Fig. 4c–f). Consistent with the hepatic and peripheral insulin-sensitizing effects of rFGF1, insulin-stimulated AKT signalling was enhanced in both the liver and muscle of chronically rFGF1-treated DIO mice (Extended Data Fig. 4g, h). Taken together, these findings demonstrate that chronic administration of rFGF1 leads to sustained glucose lowering and whole-body insulin sensitization.

The above studies demonstrating robust and sustained glucose lowering in multiple diabetic mouse models raised the possibility of parenteral administration of rFGF1 as a diabetic therapy. In support of this notion, diabetic mice chronically treated with rFGF1 were phenotypically normal in terms of locomotor activity, oxygen consumption and respiratory exchange ratio (Extended Data Fig. 4i–n), suggesting that chronic treatment with rFGF1 does not evoke adverse pleiotropic effects. Although endogenous FGF1 has been shown to have a role in adipose remodelling, histological examination of this tissue found no abnormalities after chronic treatment with rFGF1 (Extended Data Fig. 4o). Furthermore, serum creatine kinase levels did not change in chronically rFGF1-treated *ob*/*ob* mice, indicating the absence of muscle tissue damage (Extended Data Fig. 4p).

Although thiazolidinediones are the only therapeutic insulin sensitizers, their application is limited by adverse side effects, including weight gain, increased liver steatosis and bone fractures[5]. Notably, chronic treatment with rFGF1 did not lead to weight gain, and there was a decrease in hepatic steatosis (Fig. 2f, Extended Data Fig. 2i, j and Extended Data Table 1). Similar to thiazolidinediones, endocrine FGF21, which has recently been shown to have potential therapeutic blood glucose-lowering effects, has also been associated with a loss of bone density[14]. In contrast, bone mineral density, trabecular bone architecture and cortical bone thickness were not affected by chronic treatment with rFGF1 in DIO mice, as determined by micro-computed tomography analyses (Extended Data Fig. 4q, r). Furthermore, total and high-molecular-weight serum adiponectin levels were not altered by chronic treatment with rFGF1, differentiating its mechanism of action from the adiponectin-dependent glucose-lowering effects of FGF21 (refs 15, 16) (Extended Data Fig. 4s). As the prototype of a growth factor family, rFGF1 has the potential to induce unwanted cell proliferation, and concern resides in whether the mitogen properties of rFGF1 could be dissociated from its glucose-lowering activities. To address this question we generated an FGF1 ligand, rFGF1$^{\Delta NT}$, lacking the first 24 residues from the amino terminus. On the basis of the crystal structures of FGF1–FGFR complexes, the truncation was predicted to have significant effects on the binding affinity of FGF1 for all FGFRs, and similar truncations have been shown to decrease FGF1 mitogenicity. Consistent with this prediction, rFGF1$^{\Delta NT}$ showed a marked decrease in binding affinity for FGFRs compared with the native ligand, yet was still able to bind FGFR1c and FGFR2c, although with lower affinity (Extended Data Fig. 5a, b). *In vitro*, rFGF1$^{\Delta NT}$ showed a somewhat attenuated ability to activate intracellular signals downstream of FGFRs but showed a severe decrease in mitogenic activity (Fig. 4d and Extended Data Fig. 6a). Parenteral delivery of rFGF1$^{\Delta NT}$ dose-dependently lowered blood glucose levels in both genetically induced and diet-induced mouse models of diabetes (Fig. 4e, f and Extended Data Fig. 6b). rFGF1$^{\Delta NT}$ also retained the feeding suppression effects observed with rFGF1 (Extended Data Fig. 6c). The synthetic effects of exogenous rFGF1 on physiology, such as glucose homeostasis and feeding behaviour, therefore differ from, and are independent of, its classical role as a growth factor and mitogen.

In exploring the receptor dependence of the observed glucose-lowering effects, we speculated that the effects were mediated through FGFR1, on the basis of its known role in insulin sensitivity[10] and the observation that rFGF1$^{\Delta NT}$ retained FGFR1c-binding affinity in our surface plasmon resonance studies. Furthermore, our previous studies on *Fgf1* knockout mice implicated adipose tissue as a major site of FGF1 action. Accordingly, we generated mice lacking *Fgfr1* predominantly in adipose tissue (aP2-Cre;*Fgfr1*$^{fl/fl}$, R1 knockout mice). Indeed, although rFGF1 potently lowered blood glucose levels in control diabetic mice, it failed to lower glucose levels in the R1 knockout mice (Fig. 4g), indicating a requirement for FGFR1. Consistent with this, rFGF1$^{\Delta NT}$ similarly failed to affect blood glucose levels in R1 knockout mice (Extended Data Fig. 6d). An FGF1 analogue with severely attenuated FGFR-mediated signalling *in vitro* (FGF1 L29–D155; rFGF1$^{\Delta NT2}$) did not significantly affect blood glucose levels in diabetic mice (Extended Data Fig. 6e, f), further supporting

the notion that FGFR1-mediated signalling is required for the glucose-lowering effects of parenteral rFGF1. Taken together, these studies identify exogenous or endocrinized rFGF1 as a potent and long-lasting insulin sensitizer that seems to circumvent the adverse side effects associated with other diabetic therapies.

Previously we identified a role for endogenous FGF1 in the adaptive remodelling of visceral adipose tissue in response to nutrient fluctuations. The profound metabolic dysregulation observed in FGF1 knockout mice when stressed by a high-fat diet was associated with decreased vascularity in visceral adipose depots, consistent with the known role of FGF1 in angiogenesis. In contrast, our present findings identify a neomorphic insulin-sensitizing action for FGF1 in which systemic delivery of the normally autocrine/paracrine FGF1 through parenteral routes results in potent and sustained correction of hyperglycaemia accompanied by whole-body insulin sensitization.

The apparent divergent activities *in vivo* now ascribed to FGF1 seem to closely parallel those of FGF21, which has been assigned locally restricted roles in adipose tissue as well as systemic glucose-lowering activities. However, FGF21 circulates as a true endocrine hormone, whereas the high-affinity heparan sulphate proteoglycan-binding activity and serum lability of FGF1 restrict its endogenous actions to local tissues, resulting in the rapid clearance of exogenous FGF1 from the circulation[1]. In addition, these two FGFs have disparate FGFR specificities: whereas FGF1 can bind and signal through each of the alternatively spliced forms of the FGFRs, FGF21 signalling requires a heterodimeric β-klotho–FGFR complex[7]. Our findings indicate that the metabolic effects of exogenous FGF1 are mediated through FGFR1 in adipose tissue; however, additional studies will be necessary to exclude the involvement of additional receptors and/or tissues in the body-wide effects attributed to parenterally delivered rFGF1. In conclusion, given that the glucose-lowering effects of rFGF1 were not accompanied by side effects associated with current insulin sensitizers, along with the discovery that non-mitogenic rFGF1$^{\Delta NT}$ retains glucose-normalizing capability, rFGF1 and its derivatives may hold therapeutic promise.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 7 August 2013; accepted 29 May 2014.
Published online 16 July 2014.

1. Beenken, A. & Mohammadi, M. The FGF family: biology, pathophysiology and therapy. *Nature Rev. Drug Discov.* **8,** 235–253 (2009).
2. Itoh, N. & Ornitz, D. M. Fibroblast growth factors: from molecular evolution to roles in development, metabolism and disease. *J. Biochem.* **149,** 121–130 (2011).
3. Jonker, J. W. *et al.* A PPARγ–FGF1 axis is required for adaptive adipose remodelling and metabolic homeostasis. *Nature* **485,** 391–394 (2012).
4. Sun, K. & Scherer, P. E. The PPARγ–FGF1 axis: an unexpected mediator of adipose tissue homeostasis. *Cell Res.* **22,** 1416–1418 (2012).
5. Lehrke, M. & Lazar, M. A. The many faces of PPARγ. *Cell* **123,** 993–999 (2005).
6. Kharitonenkov, A. *et al.* FGF-21 as a novel metabolic regulator. *J. Clin. Invest.* **115,** 1627–1635 (2005).
7. Dutchak, P. A. *et al.* Fibroblast growth factor-21 regulates PPARγ activity and the antidiabetic actions of thiazolidinediones. *Cell* **148,** 556–567 (2012).
8. Zinn, K. R. *et al.* Imaging Tc-99m-labeled FGF-1 targeting in rats. *Nucl. Med. Biol.* **27,** 407–414 (2000).
9. Lee, J. & Blaber, M. The interaction between thermodynamic stability and buried free cysteines in regulating the functional half-life of fibroblast growth factor-1. *J. Mol. Biol.* **393,** 113–127 (2009).
10. Wu, A. L. *et al.* Amelioration of type 2 diabetes by antibody-mediated activation of fibroblast growth factor receptor 1. *Sci. Transl. Med.* **3,** 113ra126 (2011).
11. Li, A. J., Tsuboyama, H., Komi, A., Ikekita, M. & Imamura, T. Strong suppression of feeding by a peptide containing both the nuclear localization sequence of fibroblast growth factor-1 and a cell membrane-permeable sequence. *Neurosci. Lett.* **255,** 41–44 (1998).
12. Suzuki, S. *et al.* Feeding suppression by fibroblast growth factor-1 is accompanied by selective induction of heat shock protein 27 in hypothalamic astrocytes. *Eur. J. Neurosci.* **13,** 2299–2308 (2001).
13. Sasaki, K. *et al.* Effects of fibroblast growth factors and related peptides on food intake by rats. *Physiol. Behav.* **56,** 211–218 (1994).
14. Wei, W. *et al.* Fibroblast growth factor 21 promotes bone loss by potentiating the effects of peroxisome proliferator-activated receptor γ. *Proc. Natl Acad. Sci. USA* **109,** 3143–3148 (2012).
15. Holland, W. L. *et al.* An FGF21-adiponectin-ceramide axis controls energy expenditure and insulin action in mice. *Cell Metab.* **17,** 790–797 (2013).
16. Lin, Z. *et al.* Adiponectin mediates the metabolic effects of FGF21 on glucose homeostasis and insulin sensitivity in mice. *Cell Metab.* **17,** 779–789 (2013).

Acknowledgements We thank L. Chong, J. Alvarez, S. Kaufman, B. Collins, X. Zhao, S. Liu, A. Jurdzinski, A. Bleeker, K. Bijsterveld, D. Oh and G. Bandyopadhyay for technical assistance, and L. Ong and C. Brondos for administrative assistance. Computed tomography was performed at the Veterans Medical Research Foundation. R.M.E. is a Howard Hughes Medical Institute Investigator at the Salk Institute and March of Dimes Chair, and is supported by National Institutes of Health (NIH) grants (DK057978, DK090962, HL088093, HL105278 and ES010337), the Glenn Foundation for Medical Research, the Leona M. and Harry B. Helmsley Charitable Trust, Ipsen/Biomeasure, the California Institute for Regenerative Medicine and The Ellison Medical Foundation. C.L. and M.D. are funded by the National Health and Medical Research Council (grants 512354, 632886 and 1043199); J.W.J. by the European Research Council (grant IRG-277169), the Human Frontier Science Program (CDA00013/2011-C), the Netherlands Organisation for Scientific Research (VIDI grant 016.126.338), the Dutch Digestive Foundation (grant WO 11-67) and the Dutch Diabetes Foundation (grant 2012.00.1537); J.M.O. by NIH grants (DK-033651, DK-074868, T32-DK-007494, DK-063491 and P01-DK054441-14A1) and the Eunice Kennedy Shriver National Institute of Child Health and Human Development/NIH through cooperative agreement of U54-HD-012303-25 as part of the specialized Cooperative Centers Program in Reproduction and Infertility Research; M.M. by the National Institute of Dental and Craniofacial Research grant (DE13686); and M.A. by an F32 Ruth L. Kirschstein National Research Service Award (National Institute of Diabetes and Digestive and Kidney Diseases).

Author Contributions J.M.S., J.W.J. M.D. and R.M.E. designed and supervised the research. J.M.S., J.W.J., M.A., R.G., D.L., O.O., Z.H., W.L., E.Y., T.H.D., R.H., W.F., Y.-Q.Y. and A.R.A. performed research. J.M.S., J.W.J., M.A., R.T.Y., C.L., A.R.A., J.M.O., M.M., M.D. and R.M.E. analysed data. J.M.S., J.W.J., M.A., R.G., A.R.A., M.D. and R.M.E. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.M.E. (evans@salk.edu) or M.D. (downes@salk.edu).

LETTER

doi:10.1038/nature13492

Coordinated regulation of protein synthesis and degradation by mTORC1

Yinan Zhang[1], Justin Nicholatos[1], John R. Dreier[2], Stéphane J. H. Ricoult[1], Scott B. Widenmaier[1], Gökhan S. Hotamisligil[1], David J. Kwiatkowski[2] & Brendan D. Manning[1]

Eukaryotic cells coordinately control anabolic and catabolic processes to maintain cell and tissue homeostasis. Mechanistic target of rapamycin complex 1 (mTORC1) promotes nutrient-consuming anabolic processes, such as protein synthesis[1]. Here we show that as well as increasing protein synthesis, mTORC1 activation in mouse and human cells also promotes an increased capacity for protein degradation. Cells with activated mTORC1 exhibited elevated levels of intact and active proteasomes through a global increase in the expression of genes encoding proteasome subunits. The increase in proteasome gene expression, cellular proteasome content, and rates of protein turnover downstream of mTORC1 were all dependent on induction of the transcription factor nuclear factor erythroid-derived 2-related factor 1 (NRF1; also known as NFE2L1). Genetic activation of mTORC1 through loss of the tuberous sclerosis complex tumour suppressors, TSC1 or TSC2, or physiological activation of mTORC1 in response to growth factors or feeding resulted in increased NRF1 expression in cells and tissues. We find that this NRF1-dependent elevation in proteasome levels serves to increase the intracellular pool of amino acids, which thereby influences rates of new protein synthesis. Therefore, mTORC1 signalling increases the efficiency of proteasome-mediated protein degradation for both quality control and as a mechanism to supply substrate for sustained protein synthesis.

In response to growth signals, mTORC1 promotes anabolic processes, such as protein synthesis, and its chronic activation is believed to underlie a variety of complex human diseases, including cancer and metabolic diseases, as well as pathologies associated with ageing[1,2]. Cells must possess mechanisms to coordinate protein synthesis with protein turnover to maintain amino acid and protein homeostasis, as even a small persistent imbalance between these processes can disrupt cell and tissue physiology[3,4]. Given the ubiquitous role of mTORC1 in stimulating protein synthesis, we sought to assess the effects of mTORC1 activation on protein degradation.

A protein complex comprising TSC1, TSC2 and TBC1D7 (the TSC complex) serves as a central negative regulator of mTORC1, with loss of its components resulting in growth-factor-independent activation of mTORC1 (ref. 1). *Tsc2*$^{-/-}$ mouse embryonic fibroblasts (MEFs) exhibit growth-factor-independent activation of mTORC1, as scored by phosphorylation of S6K1 and S6 (Extended Data Fig. 1a), which is blocked by the mTORC1 inhibitor rapamycin or reconstitution with human *TSC2*. Both TSC1- and TSC2-deficient MEFs displayed a 20–25% increase in the rate of *de novo* protein synthesis, which was abolished by rapamycin (Fig. 1a and Extended Data Fig. 1b). To analyse relative turnover rates of newly synthesized proteins, the percentage of total labelled protein remaining over time was measured in a pulse-chase experiment (Extended Data Fig. 1c). Both TSC1- and TSC2-deficient cells displayed a rapamycin-sensitive increase in the rate of protein degradation (Fig. 1b and Extended Data Fig. 1d–f). This was surprising given the well-established role of mTORC1 in inhibiting autophagy, a lysosome-dependent mechanism of degrading organelles and proteins[5]. While the lysosome inhibitor chloroquine slowed the rate of protein degradation, cells lacking TSC2 maintained a rapamycin-sensitive increase in protein turnover, and rapamycin also slowed rates of protein degradation in autophagy-deficient (*Atg7*$^{-/-}$)[6] MEFs (Fig. 1c and Extended Data Fig. 2a–c).

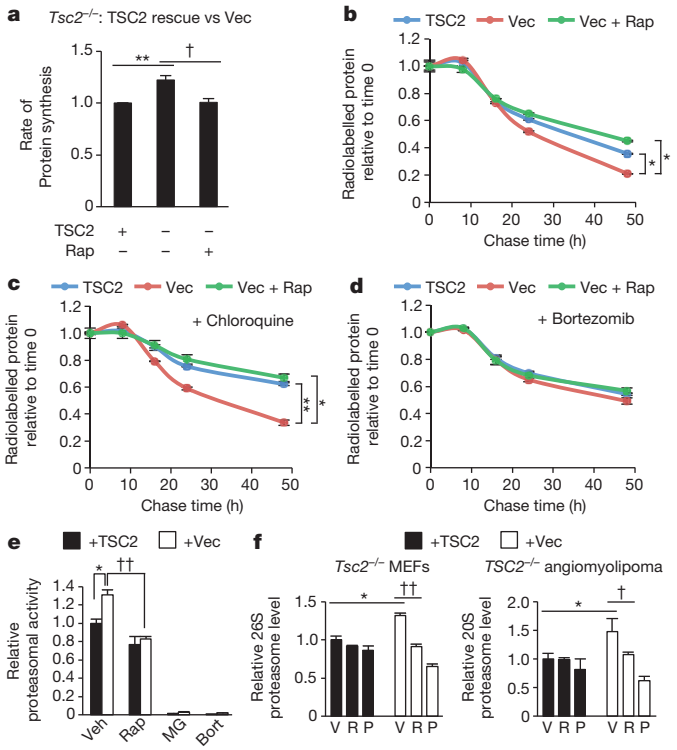


Figure 1 | mTORC1 enhances protein degradation through an increase in proteasome levels. **a**, *Tsc2*$^{-/-}$ MEFs expressing TSC2 or empty vector (Vec) were serum starved for 16 h with vehicle or 20 nM rapamycin (Rap), and protein synthesis was measured with ^{35}S-Met incorporation (20 min). Data are mean ± standard error of the mean (s.e.m.) ($n = 3$). $^{**}P < 0.01$, †$P < 0.05$. **b**, Cells treated as in **a** were pulse labelled for 30 min and chased in medium containing vehicle or rapamycin. The rate of protein degradation is shown as the fraction of radiolabelled protein remaining over time. **c**, **d**, Cells were treated as in **b**, except that 10 μM chloroquine (**c**) or 0.02 μM bortezomib (**d**) was present in the chase media. **b**–**d**, Data are mean ± s.e.m. ($n = 3$; note: small error bars are masked by line symbols). $^{*}P < 0.05$, $^{**}P < 0.01$, at 48 h. **e**, The cells from **a** were serum starved for 16 h with vehicle (Veh) or 20 nM rapamycin, or treated for 2 h with MG132 (MG; 0.1 μM) or bortezomib (Bort; 0.2 μM). Proteasome activity is presented as mean ± s.e.m. relative to vehicle-treated cells expressing TSC2 ($n = 3$). $^{*}P < 0.05$, ††$P < 0.01$. **f**, Cells were serum starved for 24 h with vehicle (V), 20 nM rapamycin (R) or 2.5 μM PP242 (P). Intact proteasome levels are presented as mean ± s.e.m. relative to vehicle-treated TSC2-expressing cells ($n = 3$). Graphs are labelled as in **e**; $^{*}P < 0.05$, †$P < 0.05$, ††$P < 0.01$. **a**–**e**, Statistical significance for pairwise comparisons evaluated with a two-tailed Student's *t*-test.

[1]Department of Genetics and Complex Diseases, Harvard School of Public Health, Boston, Massachusetts 02115, USA. [2]Translational Medicine Division, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA.

A role for the proteasome as the other major cellular mechanism of proteolysis was examined. As with rapamycin, two structurally distinct proteasome inhibitors, bortezomib and MG132, blocked the enhanced rate of protein degradation in TSC2-deficient cells (Fig. 1d and Extended Data Fig. 2d–f). TSC2-deficient cells possessed significantly higher proteasome activity relative to TSC2-expressing cells, which was attenuated by rapamycin and abolished with proteasome inhibitors (Fig. 1e). The levels of intact proteasomes were measured in four distinct isogenic pairs of cells, including $Tsc2^{-/-}$ MEFs and a $TSC2^{-/-}$ human angiomyolipoma-derived cell line expressing either empty vector or human TSC2, and HeLa and MCF10A cells stably expressing control or TSC2-targeting short hairpin (sh)RNAs. TSC2 loss and mTORC1 activation resulted in a significant increase in cellular proteasome content in all four lines, which was reversed by rapamycin or the mTOR kinase inhibitor PP242 (Fig. 1f and Extended Data Fig. 2g, h).

To determine the mechanism of proteasome increase downstream of mTORC1, we focused on transcriptional regulation, as induced expression of the genes encoding proteasome (PSM) subunits has been established as a major mechanism controlling cellular proteasome content[7,8]. Interestingly, in a previous transcriptional profiling study, gene set enrichment analysis of mTORC1-induced transcripts found that the most enriched gene set was 'Parkin disorder under Parkinson disease'[9]. In examining the genes driving this enrichment score, we found that ten PSM genes scored as being significantly stimulated by mTORC1 (Extended Data Fig. 3a). We confirmed that mTORC1 activation induced the expression of messenger RNAs encoding subunits of both the 20S core and 19S regulatory complex of the proteasome in two independent sets of *Tsc2*-null MEFs and in HeLa cells with stable knockdown of *TSC2*, as well as wild-type MEFs stimulated with serum (Fig. 2a and Extended Data Fig. 3b–e). NRF1 has been demonstrated to induce the global expression of PSM genes through direct binding of shared regulatory elements in their promoters[7,8]. Importantly, short interfering (si)RNA knockdown of NRF1, but not the closely related NRF2 (also known as NFE2L2), blocked the mTORC1-dependent induction of PSM genes and led to a decrease in intact proteasome levels in multiple TSC2-deficient cell lines (Fig. 2b, c and Extended Data Fig. 4a–d). Reciprocally, exogenous overexpression of two different *Nrf1* complementary DNA constructs led to elevated levels of intact proteasomes, which, unlike control cells, were resistant to the effects of rapamycin (Fig. 2d and Extended Data Fig. 4e). Like rapamycin, NRF1 knockdown blocked the enhanced rate of protein turnover in TSC2-deficient cells (Fig. 2e and Extended Data Fig. 4f, g). Collectively, these data indicate that NRF1 functions downstream of mTORC1 in promoting proteasome-mediated protein degradation.

A rapamycin-sensitive increase in NRF1, but not in NRF2, protein levels was observed in multiple mouse and human cell lines lacking TSC2, as well as in HEK293 cells overexpressing RHEB, the downstream target of the TSC complex that activates mTORC1 (Fig. 2b and Extended Data Figs 4b, c, f and 5a–c)[1]. In wild-type cells, mTORC1 signalling is dependent on growth factors, and NRF1, but not NRF2, was upregulated in a rapamycin-sensitive manner over a time course of growth factor stimulation (serum, epidermal growth factor (EGF) or insulin) in wild-type MEFs, MCF10A and HeLa cells (Fig. 3a and Extended Data Fig. 5b, c). The mTORC1-mediated induction of NRF1, through either growth factors or TSC2 loss, was reflected in an increase in both the unprocessed (p120) and processed (p110) isoforms of NRF1 (ref. 10), as seen most clearly on a NuPage gradient gel (Extended Data Fig. 5d). Rapamycin also decreased the protein levels of individual proteasome subunits, as well as an insulin-stimulated increase in intact 26S proteasomes (Extended Data Fig. 5e, f).

Chronic mTORC1 signalling can cause endoplasmic reticulum (ER) stress and activate the unfolded protein response (UPR)[11], an adaptive response that includes increased proteasomal degradation of ER proteins[12]. Indeed, TSC2-deficient cells displayed a rapamycin-sensitive increase in phosphorylation of the UPR effector PERK (also known as EIF2AK3) (Extended Data Fig. 6a). However, classical chemical inducers of ER stress, tunicamycin and thapsigargin, failed to induce NRF1 in wild-type cells. Consistent with previous studies[7,8], treatment of cells with proteasome inhibitors led to increased NRF1 protein levels, comparable to those seen in TSC2-deficient cells. However, in contrast to TSC2 depletion,

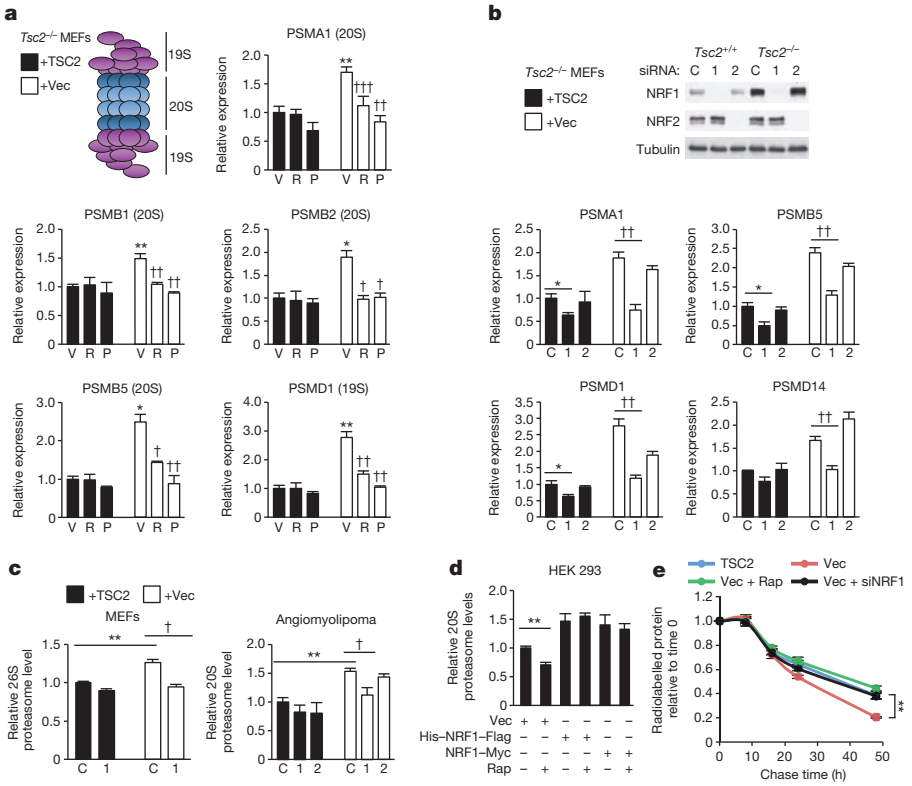


Figure 2 | mTORC1 induces proteasome gene expression and protein degradation through NRF1. a, $Tsc2^{-/-}$ MEFs expressing TSC2 or empty vector (Vec) were serum starved for 16 h with vehicle or 20 nM rapamycin. Transcript levels are shown as mean ± s.e.m. relative to vehicle-treated TSC2-expressing cells ($n = 3$). $^{*}P < 0.05$ or $^{**}P < 0.01$ compared to vehicle-treated TSC2-expressing cells; $^{\dagger}P < 0.05$, $^{\dagger\dagger}P < 0.01$, or $^{\dagger\dagger\dagger}P < 0.001$ compared to vehicle-treated vector-expressing cells. P, PP242; R, rapamycin; V, vehicle. **b**, siRNA-transfected cells were serum starved for 16 h. Transcript levels are presented as mean ± s.e.m. relative to TSC2-expressing cells with control siRNAs ($n = 3$). $^{*}P < 0.05$, $^{\dagger\dagger}P < 0.01$. C, control siRNA; 1, NRF1 siRNA; 2, NRF2 siRNA. **c**, Proteasome levels for cells treated as in **b** are presented as in **b** ($n = 3$). $^{**}P < 0.01$, $^{\dagger}P < 0.05$. **d**, Proteasome levels in HEK293 cells transfected with indicated plasmids and serum starved for 16 h with vehicle or 20 nM rapamycin (Rap) are shown as mean ± s.e.m. relative to vehicle-treated vector-expressing cells. $^{**}P < 0.01$. **e**, Rates of protein degradation in serum-starved siRNA-transfected cells (control or NRF1) treated with vehicle or rapamycin are shown as the fraction of radiolabelled protein remaining over time, presented as mean ± s.e.m. ($n = 3$; note: small error bars are masked by line symbols). $^{**}P < 0.01$ at 48 h. **a**–**e**, Statistical significance for pairwise comparisons evaluated with a two-tailed Student's *t*-test.

RHEB overexpression and growth factor stimulation, the proteasome-inhibitor-induced increase in NRF1 levels was not reversed by rapamycin. Therefore, mTORC1 signalling increases NRF1 levels in a manner that is independent of both the UPR and the proteasome recovery pathway. Finally, we failed to detect effects of mTORC1 on the nuclear and cytosolic distribution of NRF1 (Extended Data Fig. 6b, c).

The timing of NRF1 induction upon mTORC1 activation (6 to 12 h), and the fact that expression of NRF1 from an exogenous promoter led to elevated levels of NRF1 and proteasomes that were no longer sensitive to rapamycin, suggested that mTORC1 might regulate NRF1 through transcriptional control mechanisms. Indeed, *Nrf1* transcript levels were elevated and sensitive to mTOR inhibitors in *Tsc2*-null cells (Fig. 3b). Amongst the transcription factors identified to be downstream of mTORC1 (ref. 9), we found that the sterol regulatory element binding protein 1 (SREBP1; also known as SREBF1) regulated *Nrf1* gene expression. siRNA-mediated knockdown of SREBP1, but not SREBP2, decreased *Nrf1* transcript levels to a similar extent as rapamycin in both TSC2-deficient MEFs (Fig. 3c) and human angiomyolipoma-derived cells (Extended Data Fig. 7a), and SREBP1 depletion also decreased NRF1 protein levels (Fig. 3d). Reciprocally, exogenous expression of mature, active SREBP1c upregulated NRF1 and rendered its expression resistant to rapamycin (Fig. 3e). Analysis of previous genomic data sets from studies aimed at identifying targets of SREBP1, including genome-wide expression[13] and chromatin immunoprecipitation (ChIP)[14], further suggested that the *NRF1* gene is directly regulated by SREBP1. Bioinformatic analysis of the human and rodent *NRF1* loci identified four consensus and conserved sterol regulatory elements in proximity with the two predicted transcription start sites (Extended Data Fig. 7b). A ChIP assay demonstrated that mature SREBP1c bound to the *NRF1* promoter, where Pol II binding was also enriched (Fig. 3f). As controls, SREBP1c also bound to the promoter of its established target *SCD* but not to the promoters of *GAPDH* or *NRF2* (Extended Data Fig. 7c). As mTORC1 signalling increases the accumulation of active SREBP1 (ref. 9), these collective findings indicate that SREBP1 lies downstream of mTORC1 in the induction of *NRF1* expression.

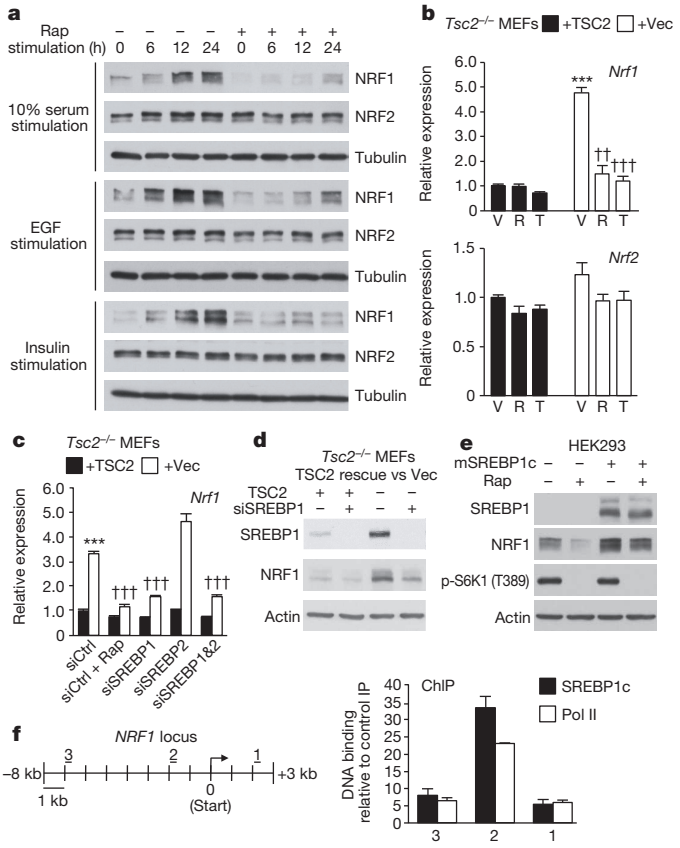


Figure 3 | Growth factors stimulate an increase in NRF1 through mTORC1, which induces *NRF1* transcription in an SREBP1-dependent manner. **a**, *Tsc2*$^{-/-}$ MEFs expressing TSC2 were serum starved for 16 h and stimulated with 10% serum, 10 ng ml^{-1} EGF or 100 nM insulin for the indicated durations with vehicle or 20 nM rapamycin (Rap). **b**, *Nrf1* and *Nrf2* transcript levels from cells that were serum starved for 16 h with vehicle (V), 20 nM rapamycin (R) or 250 nM of the mTOR inhibitor torin 1 (T) are shown as mean ± s.e.m. relative to vehicle-treated TSC2-expressing cells ($n = 3$). Vec, vector. ***$P < 0.001$ compared to vehicle-treated TSC2-expressing cells; ††$P < 0.01$ or †††$P < 0.001$ compared to vehicle-treated vector-expressing cells. **c**, siRNA-transfected cells were treated as in **b.** Rap, 20 nM rapamycin. *Nrf1* transcript levels are presented as mean ± s.e.m. relative to TSC2-expressing cells with control siRNAs (siCtrl; $n = 3$). ***$P < 0.001$ compared to TSC2-expressing cells, †††$P < 0.001$ compared to vector-expressing cells with control siRNAs. **d**, NRF1 protein levels from cells treated as in **c**. **e**, NRF1 protein levels in HEK293 cells transfected with mature Flag-tagged SREBP1c or empty vector and serum starved for 16 h with vehicle or 20 nM rapamycin. Phosphorylated (p) S6K1 is shown as a marker of mTORC1 activity. **f**, ChIP from HEK293 cells transfected as in **e** with anti-Flag (SREBP1c) or Pol II. Bound DNA was measured by polymerase chain reaction with quantitative reverse transcription (qRT–PCR) for the indicated promoter regions (left; 1, 2, 3) and normalized to control IgG immunoprecipitations (IPs). Data are mean ± s.e.m. ($n = 3$). **b**, **c**, Statistical significance for pairwise comparisons was evaluated with a two-tailed Student's *t*-test.

We next examined mouse models of both genetic and physiological activation of mTORC1 signalling in the brain and liver, respectively. A brain model of tuberous sclerosis complex involving a conditional hypomorphic allele of *Tsc2* (*Tsc2*$^{c\text{-}del3}$) was used, in which exon 3 is deleted in neurons through Cre expression from the synapsin I promoter (*Syn1-cre*)[15]. Brain lysates from *Tsc2*$^{+/+}$, *Tsc2*$^{c\text{-}del3/-}$ (heterozygotes) and *Tsc2*$^{c\text{-}del3/c\text{-}del3}$;*Syn1-cre* (neuron-specific deletion) mice were compared. This allelic series showed a graded loss of TSC2 protein with a corresponding increase in mTORC1 signalling, elevated NRF1 protein and mRNA levels, with no change in *Nrf2* expression, and a corresponding increase in PSM transcript and protein levels (Fig. 4a–c and Extended Data Fig. 8a). NRF1 has been shown to control PSM gene expression in hepatocytes[16]. mTORC1 signalling is strongly activated in the liver upon feeding[17]. This stimulation was associated with an increase in the protein and mRNA levels of NRF1, but not NRF2, as well as representative PSM transcripts, and these were blocked with a single dose of rapamycin just before feeding (Fig. 4d, e and Extended Data Fig. 8b). These data provide *in vivo* support for NRF1 activation and proteasome induction downstream of mTORC1 signalling.

We hypothesized that, in addition to serving as a quality control mechanism for newly translated proteins, the enhanced proteasome activity upon mTORC1 activation could serve to maintain adequate pools of amino acids to sustain new protein synthesis. Indeed, while inhibition of translation with rapamycin or cycloheximide increased amino acids, proteasome inhibitors significantly depleted intracellular amino acids (Fig. 4f and Extended Data Fig. 9a). Likewise, two distinct siRNAs targeting NRF1 elicited a significant decrease in intracellular amino acids, similar to bortezomib treatment (Fig. 4f and Extended Data Fig. 9b). Depletion of intracellular amino acids upon *Nrf1* knockdown or bortezomib treatment was also reflected in a decreased rate of protein synthesis, which was much more pronounced under conditions of lower exogenous amino acids (Fig. 4g, h and Extended Data Fig. 9c, d). However, the differences in protein synthesis under low and high amino acid conditions were not reflected in differences in mTORC1 signalling (Extended Data Fig. 9e). Finally, TSC2-deficient MEFs and MCF10A cells exhibited increased sensitivity to *NRF1* knockdown (Extended Data Fig. 10a, b), indicating the importance of NRF1 induction for viability in the context of mTORC1 activation. Collectively, our findings suggest a model whereby the mTORC1-stimulated expression of NRF1 and subsequent increase in cellular proteasome activity serve as a delayed, but pre-programmed, adaptive response accompanying increased protein synthesis downstream of mTORC1 (Extended Data Fig. 10c).

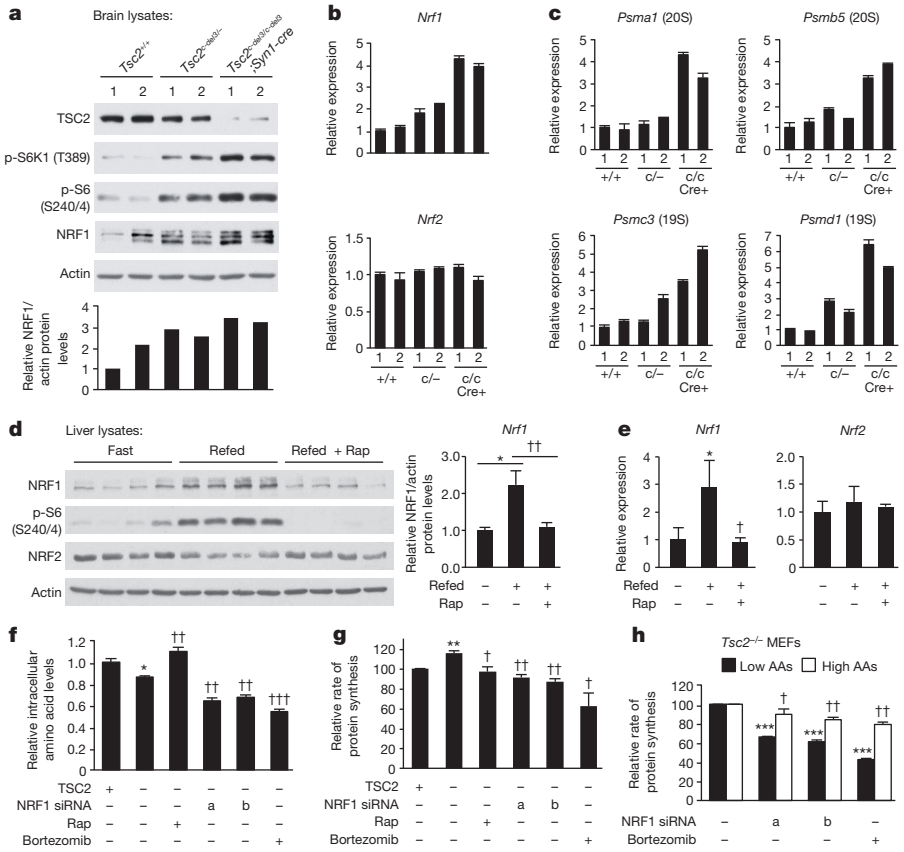


Figure 4 | NRF1 is induced upon mTORC1 activation in tissues and influences cellular amino acid levels and protein synthesis. **a**, Protein from indicated brain lysates are shown, with NRF1 levels normalized to actin below. Phosphorylated (p) S6K1 and S6 are shown as markers of mTORC1 activity. **b**, **c**, *Nrf1* and *Nrf2* (**b**) and *Psma1*, *Psmb5*, *Psmc3* and *Psmd1* (**c**) gene transcript levels from brain tissues in **a** are shown as mean ± s.e.m. of triplicate samples relative to $Tsc2^{+/+}$ sample 1. **d**, Mice fasted overnight were refed (6 h) following 30 min pre-treatment with vehicle or rapamycin (Rap; 10 mg kg^{-1}). Protein from liver lysates are shown, with NRF1 levels normalized to actin graphed as mean ± s.e.m. relative to fasted mice ($n = 4$ per condition). $^{*}P < 0.05$, $^{\dagger\dagger}P < 0.01$. **e**, Transcript levels from liver tissues in **d** are shown as mean ± s.e.m. relative to fasted mice. $^{*}P < 0.05$, $^{\dagger}P < 0.05$. **f**, $Tsc2^{-/-}$ MEFs expressing TSC2 or empty vector transfected with *Nrf1* (labelled 'a' and 'b') or control siRNAs were serum starved for 16 h with vehicle or 20 nM rapamycin or treated for 1 h with 100 nM bortezomib. Amino acid levels are shown as mean ± s.e.m. of triplicate samples relative to TSC2-expressing cells. $^{*}P < 0.05$ compared to TSC2-expressing cells, $^{\dagger\dagger}P < 0.01$ or $^{\dagger\dagger\dagger}P < 0.001$ compared to vehicle-treated vector-expressing cells. **g**, Rates of protein synthesis in cells treated as in **f** are shown as mean ± s.e.m. relative to TSC2-expressing cells ($n = 3$). $^{**}P < 0.01$ compared to TSC2-expressing cells; $^{\dagger}P < 0.05$ or $^{\dagger\dagger}P < 0.01$ compared to vehicle-treated vector-expressing cells. **h**, Cells treated as in **f** were switched to low or high amino acid (AA) media overnight, and rates of protein synthesis are shown as the mean ± s.e.m. relative to vehicle-treated cells ($n = 3$). $^{***}P < 0.001$ compared to vehicle-treated low amino acid cells; $^{\dagger}P < 0.05$ or $^{\dagger\dagger}P < 0.01$ compared to vehicle-treated high amino acid cells. **d**–**h**, Statistical significance for pairwise comparisons was evaluated with a two-tailed Student's *t*-test.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 February; accepted 15 May 2014.
Published online 13 July 2014.

1. Dibble, C. C. & Manning, B. D. Signal integration by mTORC1 coordinates nutrient input with biosynthetic output. *Nature Cell Biol.* **15,** 555–564 (2013).
2. Cornu, M., Albert, V. & Hall, M. N. mTOR in aging, metabolism, and cancer. *Curr. Opin. Genet. Dev.* **23,** 53–62 (2013).
3. Suraweera, A., Munch, C., Hanssum, A. & Bertolotti, A. Failure of amino acid homeostasis causes cell death following proteasome inhibition. *Mol. Cell* **48,** 242–253 (2012).
4. Fonseca, R., Vabulas, R. M., Hartl, F. U., Bonhoeffer, T. & Nagerl, U. V. A balance of protein synthesis and proteasome-dependent degradation determines the maintenance of LTP. *Neuron* **52,** 239–245 (2006).
5. Singh, R. & Cuervo, A. M. Autophagy in the cellular energetic balance. *Cell Metab.* **13,** 495–504 (2011).
6. Ebato, C. *et al.* Autophagy is important in islet homeostasis and compensatory increase of beta cell mass in response to high-fat diet. *Cell Metab.* **8,** 325–332 (2008).
7. Radhakrishnan, S. K. *et al.* Transcription factor Nrf1 mediates the proteasome recovery pathway after proteasome inhibition in mammalian cells. *Mol. Cell* **38,** 17–28 (2010).
8. Steffen, J., Seeger, M., Koch, A. & Kruger, E. Proteasomal degradation is transcriptionally controlled by TCF11 via an ERAD-dependent feedback loop. *Mol. Cell* **40,** 147–158 (2010).
9. Düvel, K. *et al.* Activation of a metabolic gene regulatory network downstream of mTOR complex 1. *Mol. Cell* **39,** 171–183 (2010).
10. Radhakrishnan, S. K., den Besten, W. & Deshaies, R. J. p97-dependent retrotranslocation and proteolytic processing govern formation of active Nrf1 upon proteasome inhibition. *eLife* **3,** e01856 (2014).
11. Ozcan, U. *et al.* Loss of the tuberous sclerosis complex tumor suppressors triggers the unfolded protein response to regulate insulin signaling and apoptosis. *Mol. Cell* **29,** 541–551 (2008).
12. Schröder, M. & Kaufman, R. J. The mammalian unfolded protein response. *Annu. Rev. Biochem.* **74,** 739–789 (2005).
13. Rome, S. *et al.* Microarray analyses of SREBP-1a and SREBP-1c target genes identify new regulatory pathways in muscle. *Physiol. Genomics* **34,** 327–337 (2008).
14. Reed, B. D., Charos, A. E., Szekely, A. M., Weissman, S. M. & Snyder, M. Genome-wide occupancy of SREBP1 and its partners NFY and SP1 reveals novel functional roles and combinatorial regulation of distinct classes of genes. *PLoS Genet.* **4,** e1000133 (2008).
15. Yuan, E. *et al.* Graded loss of tuberin in an allelic series of brain models of TSC correlates with survival, and biochemical, histological and behavioral features. *Hum. Mol. Genet.* **21,** 4286–4300 (2012).
16. Lee, C. S., Ho, D. V. & Chan, J. Y. Nuclear factor-erythroid 2-related factor 1 regulates expression of proteasome genes in hepatocytes and protects against endoplasmic reticulum stress and steatosis in mice. *FEBS J.* **280,** 3609–3620 (2013).
17. Yecies, J. L. *et al.* Akt stimulates hepatic SREBP1c and lipogenesis through parallel mTORC1-dependent and independent pathways. *Cell Metab.* **14,** 21–32 (2011).

Acknowledgements We thank I. Ben-Sahra and L. Yang for technical assistance. This work was supported in part by Department of Defense Tuberous Sclerosis Complex Research Program grant W81XWH-10-1-0861 (B.D.M.), National Institutes of Health grants CA122617 (B.D.M.) and CA120964 (B.D.M. and D.J.K.), the Ellison Medical Foundation (B.D.M.), National Science Foundation fellowship DGE-1144152 (S.J.H.R.), and a Canadian Institutes of Health Research fellowship (S.B.W.).

Author Contributions B.D.M. and Y.Z. designed and interpreted the experiments and wrote the manuscript. Y.Z., J.N., J.R.D., S.J.H.R. and S.B.W. performed the experiments. G.S.H. and D.J.K. provided key materials and technical guidance.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.D.M. (bmanning@hsph.harvard.edu).

CORRIGENDUM

doi:10.1038/nature13676

Corrigendum: Systems survey of endocytosis by multiparametric image analysis

Claudio Collinet, Martin Stöter, Charles R. Bradshaw, Nikolay Samusik, Jochen C. Rink, Denise Kenski, Bianca Habermann, Frank Buchholz, Robert Henschel, Matthias S. Mueller, Wolfgang E. Nagel, Eugenio Fava, Yannis Kalaidzidis & Marino Zerial

Nature **464,** 243–249 (2010); doi:10.1038/nature08779

Readers alerted us to a technical mistake (a shift of one row) made during the compilation of Supplementary Table 4, resulting in the assignment of 455 gene symbols to the wrong gene IDs. For the affected genes, the cluster group, the number of positive small interfering and endoribonuclease-prepared RNAs (si/esiRNAs) per gene and the Phenoscore values were assigned to the wrong gene symbol (although they were all assigned to the correct gene ID). All data in the online database were correct. The Supplementary Information of this Corrigendum contains the corrected Supplementary Table 4. We have also redesigned our online database (containing all images and profiles) to provide a more user-friendly interface, a visual representation of the phenotypic profiles, a facilitated gene search and integration with other public internet databases. The appropriate link on page 245 of the Article should therefore be http://endosomics.mpi-cbg.de.

Supplementary Information is available in the online version of this Corrigendum.

RETRACTION

doi:10.1038/nature13588

Retraction: Genomic organization of human transcription initiation complexes

Bryan J. Venters & B. Franklin Pugh

Nature **502,** 53–58 (2013); doi:10.1038/nature12535

We reported the presence of degenerate versions of four well known core promoter elements (BRE_u, TATA, BRE_d and INR) at most measured TFIIB binding locations found across the human genome. However, it was brought to our attention by Matthias Siebert and Johannes Söding in the accompanying Brief Communication Arising (*Nature* **511,** E11–E12, http://dx.doi.org/10.1038/nature13587; 2014) that the core-promoter-element analyses that led to this conclusion were not correctly designed. Consequently, the individual core promoter elements were not statistically validated, and therefore there is no evidence of specificity for most reported core-promoter-element locations. To the best of our knowledge, the raw and processed human TFIIB, TBP and Pol II ChIP-exo data are valid, but subject to standard false discovery considerations. We therefore retract the paper. We sincerely apologize for adverse consequences that may have arisen from the error in our analyses.

TECHNOLOGY FEATURE

WHEN DISEASE STRIKES FROM NOWHERE

When healthy parents have a child with a genetic disorder, the cause is sometimes a new mutation. Tools are emerging to meet the challenge of finding such changes.

BRAEDOSTOK/SHUTTERSTOCK

Children born with disorders not readily explained by standard tests can sometimes be diagnosed through genome sequencing and analysis.

BY VIVIEN MARX

When parents find that a child is not developing as expected, the protracted doctor visits, hospital stays and examinations only add to their distress — especially when no other family member has the condition and the standard tests on the child's blood and genes shed no light on the cause. The uncertainties, costs and anguish can be devastating to families, says Michael Friez, who directs the diagnostic laboratory at the Greenwood Genetic Center in South Carolina, a non-profit organization that analyses patients' genomes for clinicians.

Every clinical geneticist has experienced the inability to identify the cause of a child's neurodevelopmental disorder, adds Roger Stevenson, a clinical geneticist also at the centre. In the early 2000s, he began seeing a family with a toddler that had severe developmental problems, including a smaller-than-average head and intellectual disability.

It was more than a decade after their first visit before sequencing revealed that the boy had a mutation in a gene called *DYRK1A*, which is thought to have a role in brain development. The finding later helped to diagnose 16 other children in the United States and Europe who had the same symptoms — and although the condition has no cure, Stevenson saw that identifying the gene comforted the boy's parents, as did knowing that there were other children like their son.

NEW MUTATIONS

What was notable about this child's case was that it involved a *de novo* mutation — one that neither parent carries in their regular complement of DNA. *De novo* mutations can occur early in the development of the embryo. They can be in parents' gametes. Around 80% of *de novo* mutations seem to occur in the father's sperm and 20% in the mother's egg, says Joris Veltman, a geneticist at Radboud ▶

University Medical Center in Nijmegen, the Netherlands, who in July published a study of *de novo* mutations in people with intellectual disabilities[1].

Disorder-causing *de novo* mutations are hard to detect — they have to be identified among a host of other, innocuous genetic changes. A number of software-based approaches are emerging to sift through sequenced genomes in search of such mutations.

As sequencing instruments and databases of genetic information become increasingly available, tool-builders hope that their software contributions can become part of routine medical care. But sequencing and analysis are different from, say, a blood cholesterol test — samples have to be prepared for the instruments, which churn out the genome sequence in snippets that must be assembled and aligned to a reference genome, such as that curated by the Genome Reference Consortium.

The results are not perfect. A patient's genome sequence can contain errors — caused by the machine misreading a letter of DNA, for example — that must be filtered out computationally. And even then, a huge number of possibilities remains. DNA bases might differ from the reference, sequences can be inserted or deleted and the number of copies of a gene can vary. Of thousands of such changes, only one might have a role in a disorder.

"There is no single magic trick that makes our method work well."

The child's DNA is then compared with that of the parents. Again, not all differences between their genomes connect to the child's disorder. Researchers use software that includes statistical analyses to determine which changes are most likely to have a role. And the tools add information, such as published data about the links between genes and disease. These results help to create lists of genetic changes, or variants, ranked by likelihood of being linked to a disorder. But variant analysis is still an emerging science, and the software tools are still maturing. Despite this, in some cases the approach turns up a specific genetic change that is likely to be the cause of a disorder.

ASSORTED VARIANTS

Finding the probable genetic culprit does not mean a treatment is available. But such results help parents to cope with the situation, says Donald Conrad, a geneticist at Washington University in St Louis, Missouri. The results also inform parents about the risk that the condition might recur in their family and help them to plan future pregnancies. And some prospective parents might opt for genetic analysis as part of *in vitro* fertilization.

Most newborns carry about 60–100 *de novo* variants, says Conrad — few of which cause any discernible problem. Software helps to sort these variants out. Conrad has developed DeNovoGear, which does statistical analysis to distinguish potentially important signals from background noise caused by experimental error[2]. The software also analyses the nature and frequency of sequencing errors. It then compares the genomes of parents, children and other family members to distinguish true *de novo* mutations from other types of genetic variation.

To improve the odds of finding such mutations, the analysis takes into account the frequency of known variation at a given site in the genome. It does so by drawing on data from the 1000 Genomes Project, an international research consortium that catalogues human genetic variation. "There is no single magic trick that makes our method work well," Conrad says. "It is just the accumulation of many different attempts to squeeze out as much information as possible."

RAISING THE ODDS

The software must contend with the errors made by the sequencing instruments — reporting a 'C' as a 'T', for example. These mistakes are rare but hard to predict, says Conrad, and may explain many false-positive results in searches for *de novo* mutations. High-throughput sequencing instruments are more prone to error in some DNA regions — which also turn out to be where cells are more likely to make mistakes when copying and repairing the genome. These tricky places account for about 15% of the genome, Conrad notes, so current methods can reliably detect *de novo* mutations only in the other 85%.

Even the best software tools come up with 2–3 times as many false positives as true positives when analysing whole-genome sequence. True positives have to be teased out with follow-up experiments — for example, by using the laborious but precise Sanger sequencing method to look at the genetic region in question. "Each sequencing platform has its own idiosyncrasies," Conrad says, and the optimal method for detecting *de novo* mutations needs to incorporate the machine's quirks into its statistical models.

Conrad is also developing statistical methods that take account of the frequency of various sequencing errors in different regions of the genome. Other software tools typically apply the same error estimates at all genomic sites. Other researchers are pursuing their own approaches.

For the study published in July[1], Veltman and his colleagues sequenced and analysed the genomes of 50 people with severe intellectual disabilities (see 'Better diagnosis'). Working with Complete Genomics (CG) in Mountain View, California, a division of the genomics giant BGI in Shenzen, China, they identified *de novo* mutations by drawing on a number of resources. They used BGI's

OMICIA

Martin Reese, chief scientific officer of Omicia.

technology and software to analyse and compare genomes and whittle down the number of possible disease-causing candidates. They did not analyse the data with other tools such as DeNovoGear, so Veltman cannot compare the methods. But the advantage with BGI's analysis suite is that the software has been matched to the specifications of the sequencing technology, he says.

All the patients in the study had previously undergone extensive testing. Protein-coding regions of their genomes had been analysed, and microarrays were used to analyse variations in gene-copy number, which can occur from person to person and also in some disorders. Veltman says that the software showed high sensitivity in detecting *de novo* mutations, which enabled a more accurate diagnosis of almost half of the patients.

INTERPRET CAREFULLY

In Veltman's view, interpreting mutations is now more possible for disorders such as intellectual disability than for diseases such as cancer or diabetes, because many cases of severe intellectual disability seem to be caused by a single mutation. But, he says, the few hundred genes that the scientific community has found to be implicated in intellectual disability form a still-incomplete list.

Veltman stresses that the sequencing quality in the study was good, but says that even the best sequencing technology can miss or misidentify *de novo* mutations. To minimize errors, researchers need to seek out the highest-quality genome sequencing, he says. Beyond that, interpreting the many genetic variations that turn up when comparing genomes — and figuring out which ones are related to a disorder — is the field's major bottleneck. Researchers also need to find better ways to analyse *de novo* mutations in the genome's non-coding regions, which are still difficult to interpret.

One suite of tools to analyse protein-coding

and non-coding genome regions is FastQForward, which integrates the software programs VAAST[3], pVAAST[4] and Phevor[5]. These tools were co-developed by Mark Yandell, a computational geneticist at the University of Utah in Salt Lake City who directs software development and computational analysis related to the Utah Genome Project. That project combines family histories from the Utah Population Database with medical records, which increasingly include DNA sequence information. The project includes family histories for more than 7 million people and medical records for around 4 million of them.

Yandell and his team are using pVAAST to analyse family pedigrees in which there is a higher frequency of disease. pVAAST searches through many genomes in parallel to find alterations. The program addresses the statistical challenge presented by genomes from people who are related, he says. And it detects *de novo* mutations.

Printed out, the large family pedigrees in the Utah database can span almost 2 metres. The ones Yandell is studying include multiple family members that have mental-health issues such as schizophrenia or depression. Mental illness has a large environmental component, but he hopes that these records can help to uncover genetic factors, he says.

Studying families might offer advantages over the more typical 'cohort analyses' of unrelated people with similar conditions. In such cohorts, the causes of mental-health problems might be quite diverse. Yandell hopes that restricting the search to extended families will make it easier to identify gene variants involved.

VAAST uses a similar approach to that of BLAST, a widely used search tool in genetics research[6]. With BLAST, a scientist can take a genetic sequence and search through many genomes to find high-probability matches to it. Similarly, VAAST compares variants in a person's genome to those collected in the 1000 Genomes Project. This comparison helps to determine the probability that a variant is causing a disease.

pVAAST extends VAAST's capabilities to family-based sequence data. Yandell also uses Phevor, which taps into resources such as the Human Phenotype Ontology, which catalogues links between gene function and human disease symptoms.

Phevor helped clinicians to diagnose a 12-year-old boy who had life-threatening diarrhoea and intestinal inflammation. Genetic analysis with VAAST had come up empty. By combining the analysis with Phevor, the researchers traced the boy's illness to a *de novo* mutation in *STAT1*, a gene involved in many intestinal disorders. The finding, which was confirmed with Sanger sequencing, enabled the boy's doctors to properly treat and stabilize his condition.

Yandell hopes that genetic analysis will soon be a routine part of clinical diagnosis. Towards that aim, he and Martin Reese, a co-developer of VAAST, developed Opal, a platform that helps clinicians to interpret and use the results from software-based genetic analyses. Reese is chief scientific officer of Omicia, a company in Oakland, California, that offers genetic analysis using several tools, including Opal, VAAST, pVAAST and Phevor.

Reese says that his company tries to fill the gap between tools developed in academia and the needs of clinicians. The VAAST algorithm does the hard-core maths to analyse the matches, score their probabilities and create a ranking, he says. The Opal software then searches for clinical and biological data about the candidate genes — added information that can help to determine which candidates are more likely to be causing the disease.

In June, Omicia began working with Laboratory Corporation of America, a large medical-testing company based in Burlington, North Carolina. Omicia will interpret genomic data as part of clinical trials.

Data analysis is Omicia's specialty. Unlike many other companies in the field, it does not do sequencing. "We're slicing and dicing the genome based on your clinical question," says Reese. His team first assesses the sequence quality and filters out typical sequencing errors before hunting for changes such as *de novo* mutations.

FUTURE MEDICINE

Eventually, clinical standards in this area will emerge, but for now service providers use the approaches they deem to be best for these complex analyses. Reese believes that many diseases, if not all of them, have contributions from *de novo* mutations. These contributions are hard to identify, he says, but whole-genome analysis raises the probability of finding them, as Veltman's study shows.

Conrad says that detection of *de novo* mutations can be a standard medical test only when the genetic complexities of diseases they cause are better understood and tool developers have found ways to address them, and after the technical issues related to high-throughput sequencing have been resolved.

Between 20% and 90% of the *de novo* mutations detected by software and with the help of whole-genome sequencing can be false positives. "Researchers can accommodate this with extensive follow-up validation experiments, but this is just simply not practical for a routine diagnostic test," Conrad says.

"We're slicing and dicing the genome based on your clinical question."

Better approaches are also needed for the tough-to-sequence regions of the genome, and the software has to cover the spectrum of mutations, from single-base changes to insertions or deletions. And researchers need to better understand changes such as large copy-number variations, regions of repetitive sequence and other types of DNA rearrangements, says Conrad.

Greenwood Genetic Center, which uses genetic analysis to diagnose patients, does its own analysis and uses commercial services. Scientists and companies doing genetic analysis will soon have access to many of the same shared resources, and Friez says that he looks forward to seeing how that will help patients with neurodevelopmental disabilities. For now, patients, their families and clinicians all face the same issue: researchers' ability to identify mutations associated with disorders is not always matched by a medical understanding of these mutations, and therapies that might arise from knowing about them are far in the future.

But genetics does deliver some answers for these patients and families, says Veltman. "From what I hear from my clinical colleagues, these families are very happy to finally get an answer — it often means closure for them, they can give the disorder in their child a place and better accept it," he says. "In regards to therapy and treatment, unfortunately options are still quite limited, but progress is being made." ■

CASE STUDY

Better diagnosis

In a study published in July[1], a team of researchers in the United States and the Netherlands analysed the genomes of 50 people who had intellectual disabilities that could not be explained through standard tests. DNA sequencing showed that each person had more than 4 million single-base variations and more than 250 copy-number variations. Computational analysis and comparison of the patients' genomes with those of their parents whittled the number down to some 60 *de novo* mutations per person, of which 1–2 were in genes identified by the software as having a connection to intellectual disability. The genomic analysis led to a more accurate diagnosis for 20 of the 50 patients.

Vivien Marx *is technology editor for* Nature *and* Nature Methods.

1. Gilissen, C. *et al. Nature* **511,** 344–347 (2014).
2. Ramu, A. *et al. Nature Methods* **10,** 985–987 (2013).
3. Yandell, M. *et al. Genome Res.* **21,** 1529–1542 (2011).
4. Hu, H. *et al. Nature Biotechnol.* **32,** 663–669 (2014).
5. Singleton, M. V. *et al. Am. J. Hum. Genet.* **94,** 599–610 (2014).
6. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. *J. Mol. Biol.* **215,** 403–410 (1990).

CAREERS

GAINING TRACTION Europe's universities are adopting tenure to draw recruits **p.451**

FUNDING US research and development still suffering from budget battles **p.451**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

FANATIC STUDIO/GETTY

EARLY-CAREER FUNDING

Big introductions

Europe's Starting Grants are ideal for young researchers with big ideas and what it takes to bring them to life.

BY QUIRIN SCHIERMEIER

On a mid-April afternoon in 2012, Francesco Ricci boarded a plane in Rome in a highly anxious state — but not because he was afraid of flying. He knew that the ten-minute presentation he was giving the next day in Brussels could propel his career into orbit. All he had to do was to persuade a panel from the European Research Council (ERC) to fund his application for a Starting Grant.

A postdoc in the chemistry department of the University of Rome Tor Vergata, Ricci had spent the greater part of two months working on his proposal to develop DNA-based nanodevices for diagnosis and treatment of cancer. After he was invited to Brussels, he prepared his talk and rehearsed it time and again in front of colleagues, encouraging them to ask the trickiest questions they could think of. He knew that his chances of success at winning a grant hinged on him giving an effective presentation.

In the hall the next day, where a dozen hopeful applicants waited to be called, a tense silence prevailed. "You try to figure out if the others are better scientists than you and what their projects might be about," Ricci says. "You know that it's possible that fewer than half of the people in the room will get a grant."

The chance to win a €1.5-million (US$1.9-million) Starting Grant is a thrilling prospect for young scientists — and for a postdoc in Italy, where funding and career opportunities for young scientists are notoriously poor, it can be the chance of a lifetime. An ERC grant — whether a Starting Grant or a Consolidator Grant for researchers at more advanced stages — means a good couple of years during which winners do not need to worry too much about funding. That is a rare luxury in today's research environment, and one that can empower research output and scientific reputation.

Launched in 2007 as the European equivalent of the US National Science Foundation, the ERC has quickly turned into the flagship funding programmes of the European Union (EU). More than €13 billion is earmarked for the agency in the EU's €80-billion Horizon 2020 research programme — the region's biggest science-funding and innovation programme ever, launched in January. The Starting Grant programme, aimed at promising EU-based early-career researchers of any nationality, appeals to young scientists because the application rules are straightforward and it favours basic science. But it is highly selective, and success demands time, effort, rigorous preparation — and, perhaps, a bit of luck.

More than 4,500 ERC grants — including 2,330 Starting Grants — have been awarded since 2007. Typically, the ERC invites two to three times as many applicants as will receive funding to give a presentation. The success rate rose from around 3% in the first call to about 15% in 2010, but fell back down to just 9% last year, owing to a sharp increase in the number of applications.

This year, almost 3,300 scientists applied, and around 330 will be selected for funding. The next call opens on 7 October and will close on 3 February 2015 (deadlines are strictly enforced). The €430 million reserved for this call — about one-quarter of ▶

the ERC's budget for the year — will be distributed among roughly 330 people. The amount to be disbursed is slightly larger than last December's initial estimate of €411 million because it expects an equally high number of applications next year and it aims to keep the success rate consistent from year to year.

Ricci's 10-minute presentation went well — and so, he felt, did the subsequent question-and-answer session with panel members. Yet he did not make the cut: an e-mail in July told him that his project had not been recommended for funding.

He was disappointed but did not give up. The reviewers' comments were not altogether discouraging, so the following year he submitted a slightly altered proposal that emphasized the medical potential of his research. That May, he returned to Brussels. He was even more nervous than in the previous year: it was his last chance, because applicants can reapply only once. This time, he felt that the presentation and subsequent question-and-answer period went less smoothly, but his persistence was rewarded: after an anxious two-month wait, he learned that his project had been funded.

OPEN TO ALL

The ERC supports researchers, irrespective of nationality, age and gender, who will be employed at, or affiliated with, host institutions in the EU or associated countries (including Iceland, Israel, Norway and Turkey) to conduct their research. The Starting Grant is one of three that it offers; the others are consolidator grants for researchers a bit further along their career path and advanced grants for eminent senior scientists. Decisions are based on the scientific merits of the applicant and the ambition and feasibility of the project.

Applicants for starting grants must have completed their PhD or equivalent 2–7 years before the call (time taken off for maternity or paternity leave, clinical training, long-term illness or national service are not part of the count) and must have published at least one paper without co-authorship of their PhD supervisor. There is no co-financing required from host institutions, and award recipients do not have to team up with other groups or companies, as in most other EU research programmes. There are also no thematic priorities: scientists in any field, including the social sciences and humanities, are equally eligible for funding.

The foremost criterion, says José Labastida, the ERC's science director, is a unique research idea that has the potential to substantially advance the knowledge in a given field. Reviewers also want to know that a candidate has what it takes to get the work done. "Apply for an ERC grant only if you have a really new and ambitious idea and you can demonstrate that you have the potential to pursue it with a team of your own," Labastida says. "A proposal that smells of incremental research, or just more of the same stuff that you have been doing before, is not going to fly."

It can be challenging for scientists in the early stages of their career to start thinking that big. Taking on full ownership of a project — for many, a completely new experience — is no easy task.

Candidates should also make sure that their CV and summary — submitted along with a full project description — meet the ERC criteria (see 'The nuts and bolts of your application'). During the first judging stage, reviewers will look only at these components. "The majority of reviewers on the panel will not be experts in your specific line of research — so send the key message to a wide audience and leave the details to the full proposal," says Labastida. The summary should outline the idea and explain why the research is important and scientifically feasible in no more than five pages. When possible, applicants should provide preliminary data. Having friends and scientists who are unfamiliar with the research read it and note what they find hard to understand will help applicants to make the message clear, he says.

The CV should highlight relevant skills and scientific achievements. It should list any stints at high-profile universities outside the applicant's home country, and previous grants or participation in EU-funded research and mobility programmes may help to show commitment and scientific potential.

DIMITRA SALMANIDOU

Chief scientists Veerle Huvenne (right) and Aggeliki Georgiopoulou plan their next expedition.

THE RIGHT TIME

Marine geoscientist Veerle Huvenne of the National Oceanography Centre in Southampton, UK, had been the chief scientist in several international research cruises before deciding in 2009 that the time was ripe to apply for a Starting Grant. She had already gathered experience in managing a collaborative marine-research project during a Marie Curie fellowship from 2005 to 2007. She wanted to use the ERC money to map coral habitats and study the biodiversity they support.

Huvenne says that advice from team members and former collaborators was extremely helpful in securing the €1.4-million grant. "Take your time to develop a really strong idea, let it mature and discuss it intensively with colleagues and peers before you start drafting a proposal," she advises. "Make sure the synopsis is rock solid. If you're invited to an interview, prepare rigorously. It's a major effort, but there is a lot to gain for your whole career."

THE RIGHT CV

The nuts and bolts of your application

To apply for a European Research Council Starting Grant, you must follow specific guidelines. Here is a summary of what you need to do — and not do.

- Your CV should be written in English and not exceed two pages.
- It should give a complete account of your academic record, including the years of your master's (if applicable) and PhD programme and the name of the university (or universities) and department (or departments) that awarded them.
- It should also include your full name and date of birth, and ought to include the URL for a current personal website, although this is not obligatory.
- It should document previous and current academic positions, fellowships and awards, teaching activities, institutional responsibilities, memberships in scientific societies and any major collaborations.
- It should clearly explain any educational or employment gaps, or unconventional career paths.
- All ongoing and submitted grants and funding can be detailed on a separate page.
- It should list only five representative publications. Any other relevant papers should be listed in the early-achievements track-record section of the application.

A model CV and application forms are available at go.nature.com/hjn9vv. Q.S.

FLORIAN RAIBLE

Kristin Tessmar-Raible (right) and PhD student Juliane Zantke work on how lunar cycles affect the activity of worms.

Compelling and intelligible writing make a difference when reviewers flooded with applications are trying to make their choice — and honest feedback from colleagues can really help, says Kristin Tessmar-Raible, a neurobiologist at the University of Vienna in Austria. Tessmar-Raible was awarded a Starting Grant in 2013 for investigating how the waxing and waning of the Moon governs animals' monthly inner clocks. The wording of the headline and sentences and the order of paragraphs and tables can be vastly improved by input from friends, colleagues and professional proofreaders, who can all offer tips on how to make a synopsis catchy and a full proposal concise and well-structured, she says. In addition, national ERC help desks in all EU countries offer grant-writing assistance, and some also offer interview-training courses.

"Every second counts, so you have to think hard about every word and every slide that you use."

Less is often more. "Avoid unnecessary information in a CV that might only conceal the things that really matter," says Ricci. Likewise, he says, presentations that are padded with data and technical details tend to be more confusing than informative. "You have only ten minutes to describe your grand vision and ambition," says Tessmar-Raible. "Every second counts, so you have to think hard about every word and every slide that you use."

Applicants should also carefully consider which of the ERC's 25 panels they would like to evaluate their proposal. "If you end up being reviewed by the 'wrong' panel it might diminish your chance of getting funded," says Erik Garnett, a physical chemist at the Institute for Atomic and Molecular Physics in Amsterdam. Garnett had moved there in 2012 from Stanford University in California with little knowledge of the funding situation in Europe. On the advice of the director of his new institute, he applied for a Starting Grant to develop nanomaterials that can be used to make high-efficiency solar cells — his first grant proposal ever — and succeeded. Before he submitted his application, he looked up the CVs of panel members of previous calls to get a feel for whether their expertise overlapped with his research. He opted for the Material and Synthesis panel because its members seemed to have more affinity for his work than did others.

Starting grants could well galvanize researchers' careers, says Huvenne. She herself is a good example: last summer, two years after her ERC-funded project launched, she was promoted from senior research fellow to team leader for the seafloor and habitat mapping group at her institution. She was surprised by how many scientists have since got seriously interested in her research.

And in Italy, universities can hire Starting-Grant winners without following the country's historically twisted routes to academic appointment. Winning an ERC grant provided Ricci a springboard to a permanent position. Twelve months after his second return from Brussels, he was promoted to associate professor. ■

Quirin Schiermeier *is a senior reporter with* Nature *in Munich, Germany.*

CAREER PROGRESSION

Europe on track

Tenure is gaining traction in Europe even as the system is slipping away in the United States, according to a study by the League of European Research Universities in Leuven, Belgium. The study surveyed 21 universities throughout Belgium, Finland, France, Germany, Italy, the Netherlands, Sweden, Switzerland, Spain and the United Kingdom. It found that seven nations are now using tenure as a way to recruit internationally and to offer researchers a clearer career path. The paper defines tenure-track as a fixed-term contract that can lead to a permanent position. Institutions surveyed in the United Kingdom, France and Spain do not have tenure systems. Meanwhile, the proportion of tenure-track positions in the United States has declined in the past 30 years, notes the study.

EDUCATION

Degrees of difference

Fewer than one-quarter of people aged 25–64 in the 34 member nations of the Organisation for Economic Co-operation and Development (OECD) earned a degree in 2012, finds an OECD report. *Education at a Glance 2014: OECD Indicators* examined education attained by adults in Europe, North America, South America and Asia. The report found that at least one-third of adults aged 25–64 in the United States, Norway and Israel had earned a degree. Chile and Austria had the lowest rates at 12% and 13%, respectively. Other nations fell between these rates. The average age for completing a doctoral research programme across the member nations was 35. Korea reported the oldest age of 40; Germany the youngest at 31.

RESEARCH AND DEVELOPMENT

Falls in funding

US federal spending on scientific research and development is projected to have fallen by 4% from 2011 by the end of this year, according to a report from the US National Science Foundation (NSF) in Arlington, Virginia. The report, which collected data from the 27 US science-funding agencies, shows that spending reached US$140 billion in 2011 and is expected to slip to $134 billion this year. The 2014 total is likely to be even lower, says an NSF spokesperson, because it does not account for a 2013 across-the-board cut to discretionary spending.

EXTRACTION

Total recall.

BY REBECCA ROLAND

Gwen hated this part of the job: getting a memory extraction from a fresh witness.

The woman sat on the other side of the table inside the tent the Chicago PD had hastily erected near the shooting in the park. She was silent, stunned, covered in blood spatter. A cup of coffee sat untouched in front of her, its smell only somewhat covering the tang of blood, the sour stench of terror. Gwen glanced at the paperwork. Meghan Johnson. Young. Pretty. Meghan's hands curled around the cup. She wore a plain wedding band.

The extractor sat on the table, the steady green light indicating it was primed and ready. It looked like a mesh helmet equipped with a tiny pack that would nestle over the forehead.

Gwen cleared her throat. Mrs Johnson? No, no need to remind her of the man who should've been sitting next to her, and too formal besides. "Meghan?" she said softly.

The woman could have been a statue, she sat so still.

"Meghan," Gwen said louder.

The woman twitched, looked up. She swayed in her seat. "Yeah." Flat.

"It's time." Gwen lifted the extractor and carried her chair around the table so that she sat beside Meghan.

The other woman leaned back, eyes widening. "N-now?"

"I'm afraid so. The quicker I retrieve your memory of the crime, the better the quality. It will make all the difference when it comes time for the trial." She laid a hand on Meghan's. The other woman was frigid. "And you won't have to relive this again. I retrieve the memory, and then I erase it." The erasure was the only decent part of this whole process. Who wanted to recall being brutalized, or watching someone hurt a child, or seeing a loved one die?

"Does it make it easier?" Meghan asked. "Erasing it?"

"It helps."

"You speak from experience?"

Gwen hesitated. She'd gone through it once, as every extractor did: to know what it was like, to better sympathize with victims, she'd erased a memory of her getting a speeding ticket. But then she'd erased other things, like bits of high school when she was shunned and most conversations with her mother, who felt the overwhelming urge to criticize everyone. She understood that those things had happened, but she didn't have to carry them around like she used to, and when she spoke with her mother, or had a bad day at work,

JACEY

she consoled herself with the fact that she could just make it all disappear. Using the extractor left her with the same feeling as a couple of glasses of wine: pleasantly sleepy and numb, with all the bad feelings shoved far down where she didn't have to deal with them. "Yes."

Meghan's eyes shimmered behind unshed tears. "Do it."

Gwen had her sign the paperwork, then fitted the extractor over Meghan's head. She pulled out her tablet and keyed in the instructions. She watched every extraction to ensure it was a useful memory for prosecuting the bastards later. Then she usually erased them from her memory. She didn't want her own baggage, much less anybody else's.

In Meghan's memory, she sat beside a handsome Latino man on a blanket in the park, only a dozen feet from the stage where a band played jazz. The details were sharp, from the quality of the sound to the way Meghan recalled the two women just in front of them swaying slightly with the music, to the lights glowing yellow in the distance along the path. She was a great witness, and she had the perfect view of the gunman as he approached and opened fire. Gwen winced. No wonder the poor woman was a mess.

Then came the moment when Meghan hovered over her husband. Blood burbled from his mouth and covered a wide swathe of his white dress shirt.

He reached for Meghan. She grasped his hands. "Love … you," he gasped.

"I love you," Meghan sobbed, over and over, until his eyes turned glassy.

The extraction stopped. Gwen was amazed at how calm her hands remained as she saved the images and sounds.

Meghan shuddered violently as if chilled to her bones.

Gwen's finger hesitated over a key. "Are you ready to remove the memory?"

She hesitated.

"You don't have to see him shot ever again. Not in your memories, not in your nightmares."

Her hands closed, opened, closed in her lap. "You remove everything?"

"Everything."

"Can you leave his last words to me?"

"I can't. I'm sorry."

"I'll keep it."

"What?" Gwen nearly hit the button out of surprise.

"I'll keep it," she whispered. "Otherwise his last words will be about our stupid mortgage. I'll keep it."

"There's no need to live with that memory."

Meghan met her gaze. Her green eyes were red-rimmed and determined. "It's worth some suffering to know that he spent his last breaths letting me know that he loved me."

Gwen's hand shook as she shut off the program and removed the extractor. Nobody had ever refused an erasure before. "I'll show you to a counsellor, if you'd like."

"I would."

Gwen led Meghan out of the tent. She held her chin high, hands steady, the complete opposite of the wreck she'd been when she'd entered. Gwen passed her to a waiting counsellor, then turned the extractor over in her hands. This was normally the time she'd erase her own memories of the scene, of the witness's testimony. *So I can stay sane*, she always told herself. Nothing good ever came from bad moments.

Except, a moment of love could sneak in. Maybe she'd been erasing good along with bad all this time. She went back into the tent and packed the extractor carefully away. ■

Rebecca Roland *is the author of* Shards of History *and the short-story collection* The King of Ash and Bones, and Other Stories. *Find out more at rebeccaroland.net.*

➲ NATURE.COM
Follow Futures:
@NatureFutures
go.nature.com/mtoodm